MODELING DATA WITH CLUMPS

By

YONGYI MIN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2003

I dedicate this work to my parents and Josh.

ACKNOWLEDGMENTS

First of all I would like to express my deepest gratitude to Dr. Alan Agresti for serving as my dissertation advisor and providing me the great opportunity to work with him. I would like to thank Drs. James Booth, Ramon Little, Rongling Wu, and James Algina for serving on my committee. In addition, I thank Dr. Alan Hutson for serving on my committee before leaving the University of Florida and providing data sets for me to use in my dissertation. I also appreciate the help of Dr. Cyndi Garvan with my dissertation. I would also like to thank the faculty and students in the Department of Statistics for their help and friendship.

I thank my parents for their constant love and confidence in my success. Finally I want to thank Josh for his proofreading of my "boring" statistics dissertation, for his support and continuous encouragement.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS iv

ABSTRACT viii

CHAPTER

1 INTRODUCTION 1

1.1 Models for Semicontinuous Data 3
1.1.1 Tobit Models 3
1.1.2 Two-Part Models 5
1.1.3 Sample Selection Models 7
1.1.4 Compound Poisson Exponential Dispersion Models 9
1.1.5 Ordinal Threshold Models 11
1.1.6 Advantages and Disadvantages of Existing Approaches . . . 12
1.2 Models for Zero-Inflated Count Data 13
1.2.1 Zero-Inflated Discrete Distributions 14
1.2.2 Hurdle Models 16
1.2.3 Finite Mixture Models 17
1.2.4 Neyman Type A Distribution 18
1.2.5 Advantages and Disadvantages of Existing Approaches . . . 19
1.3 Methods for Compliance Data Analysis 20
1.4 Modeling Repeated Measurements of Zero-Clumped Data 22
1.4.1 Repeated Measurements of Semicontinuous Data 22
1.4.2 Repeated Measurements of Zero-Inflated Data 25
1.5 Outline of Dissertation 26

2 RANDOM EFFECTS MODELS FOR REPEATED MEASURES OF ZERO-INFLATED COUNT DATA 28

2.1 The Hurdle Model 29
2.1.1 Fitting the Hurdle Model 29
2.1.2 Special Cases 33
2.1.3 Hurdle Models vs Zero-Inflated Count Models 35
2.2 Hurdle Models with Random Effects 40
2.2.1 Model Specification 40
2.2.2 ML Model Fitting with Normal Random Effects 42
2.2.3 ML Model Fitting with A Nonparametric Approach 44
2.3 Cumulative Logit Models with Random Effects 48

2.4 An Occupational Injury Prevention Program Study 50
2.4.1 Random Effects Hurdle Models 50
2.4.2 Random Effects Cumulative Logit Models 54
2.4.3 Comparison of Models 56
2.5 Identifiability of the ZIP Model 57

3 MODELING CROSS-SECTIONAL COMPLIANCE DATA 63

3.1 A Two-Part Model 63
3.1.1 Models for the First Stage 65
3.1.2 Models for Continuous Proportions Between (0,1) 67
3.2 Mixtures of Experts Models 72
3.2.1 The ME Model 73
3.2.2 Model Fitting 75
3.2.3 Standard error estimation 79
3.2.4 Model Selection 81
3.2.5 Group Comparison 83
3.3 Single-Model Approaches 84
3.3.1 Ordinal Threshold Models 85
3.3.2 Quasi-Likelihood Method 90
3.4 Application 95
3.4.1 Without Other Covariates 95
3.4.2 With Other Covariates 105
3.4.3 Summary 113

4 MODELING REPEATED MEASURES OF COMPLIANCE DATA . . 117

4.1 Mixtures of Experts Model with Random Effects 118
4.1.1 Model Specification 118
4.1.2 Nonparametric ML Model Fitting 120
4.1.3 Standard Error Estimation 126
4.1.4 Model Selection 129
4.2 Scaled Cumulative Logit Models with Random Effects 130
4.3 Marginal Models and GEE 130
4.3.1 Generalized Estimating Equations 131
4.3.2 GEE for Repeated Ordinal Response Data 133
4.4 Mixtures of Marginal Models 135
4.4.1 Extension of GEE Method for Simplex Distributions 135
4.4.2 Mixtures of Marginal Models 139
4.4.3 Expectation-Solution Algorithm 143
4.5 Applications 146
4.5.1 Asthma Study I 146
4.5.2 Asthma Study II 156

5 CONCLUSIONS 166

5.1 Summary 166
5.2 Future Research 170

REFERENCES 173

BIOGRAPHICAL SKETCH 180

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

MODELING DATA WITH CLUMPS

By

Yongyi Min

December 2003

Chair: Alan Agresti
Major Department: Statistics

Applications in which data have clumps occur in many disciplines. In this dissertation, we develop methods for modeling a few special cases of this type of data, including repeated measures of zero-inflated count data, cross-sectional compliance data, and repeated measures of compliance data.

For count responses, the situation of excess zeros (relative to what standard models allow) often occurs in many biomedical and sociological applications. Modeling repeated measures of zero-inflated count data presents special challenges. We present two types of random effects models for repeated measurements on this type of response variable. The first model is a hurdle model with random effects, which separately handles the zero observations and the positive counts. In maximum likelihood model fitting, we consider both a normal distribution and a nonparametric approach for the random effects. We also discuss a special type of the hurdle model, which can be used to test the existence of zero-inflation. The second model is a cumulative logit model with random effects, which has the simplicity of using a single model to handle the zero-inflation problem. We illustrate the proposed methods with an example from an occupational injury prevention program.

For compliance data, there are two clumps, one at 0% and one at 100%. To analyze cross-sectional compliance data, we propose a two-part model, a cumulative logit model, and a quasi-likelihood method. Our emphasis is on the mixtures of experts model (ME). We apply the EM algorithm in fitting the ME model. Then, we extend these methods into the repeated measures settings. At the subject level, we propose a random effects ME model and use a nonparametric maximum likelihood method in model fitting. At the population level, we introduce the generalized estimating equation (GEE) method and extend it to the simplex distribution. Then we combine this extension of the GEE method with the ME model to form the mixtures of marginal models. A generalization of the EM algorithm, the ES algorithm is introduced to fit the mixtures of marginal models. We use two asthma medication compliance studies for illustration of our methods.

CHAPTER 1
INTRODUCTION

Data with clumps occur in applications in many disciplines, including meteorology, economic surveys, social science and biometrics. The most typical ones are data with clumping at zero, in which data take nonnegative values but have a substantial proportion of values at zero. It includes *semicontinuous* data and *zero-inflated* count data. We refer to a variable as semicontinuous when it has a positive continuous distribution except for a probability mass at 0. Semicontinuous data are common in many areas. Examples include the amount of daily rainfall, medical or household expenditures, or concentrations of compounds. With semicontinuous data, unlike left-censored data, the zeros represent actual response outcomes. The other typical "clumped-at-zero" data are zero-inflated count data. These are data that have a higher proportion of zeros than expected under standard distributional assumptions such as the Poisson. Such data are also common in a variety of disciplines. Examples of variables that one might expect to be zero-inflated are observations for the past month of the reported number of times participating in sports activities, the number of times one has visited a doctor, and the frequency of recreational trips.

One difficulty with semicontinuous data analysis is that the existence of a probability mass at zero makes common response distributions such as the normal or gamma inappropriate for modeling the data. Likewise for zero-inflated count data, a generalized linear model based on Poisson or overdispersed count distributions usually encounters lack of fit due to disproportionately large frequencies of zeros. Thus, these types of data stimulate interesting modeling problems.

A fair amount of statistical methodology has been developed to deal with cross-sectional data with clumping at zero. For semicontinuous data, there are the *Tobit model* proposed by Tobin (1958), its various generalizations (Amemiya 1984), the compound Poisson model (Jørgensen 1987), and the two-part model (Duan, Manning, Willard, Morris, and Newhouse 1983). For zero-inflated count data, there are the zero-inflated Poisson (ZIP) model (Lambert 1992), the hurdle model (Mullahy 1986), the finite mixture model (Aitkin and Rubin 1985) and the Neyman type A distribution (Dobbie and Welsh 2001b). However, relatively little work has been done to model repeated measures of these types of data.

Besides data with clumping at zero, some applications, however, have data with two clumps, one at zero and one at the maximum possible response value. An application in which such data commonly occur is in the study of patient *compliance* (or *adherence*). Compliance is usually defined as the extent to which a subject's behavior (in terms of taking medications, following diets, or executing lifestyle changes) coincides with medical or health advice (Haynes, Taylor, and Sackett 1979). The response distribution usually has a proportion of subjects with 0% compliance, a proportion of subjects with 100% compliance, and other subjects having compliances spread between 0% and 100%. An appropriate model permits two clumps with positive probability at the extremes and treats the remaining scale as continuous. Thus far, little attention has been paid to specialized models for compliance data.

This chapter reviews methods that have been proposed for modeling data with clumping at zero and compliance data. Section 1.1 introduces models for semicontinuous data and then summarizes their advantages and disadvantages. Section 1.2 introduces models for zero-inflated count data. Existing methods for analyzing compliance data are reviewed in Section 1.3. Section 1.4 surveys extensions of models in Section 1.1 and 1.2 to handle repeated measurements, such as in longitudinal studies. The final section (Section 1.5) gives an outline of this dissertation.

1.1 Models for Semicontinuous Data

This section introduces some methods for modeling semicontinuous data. The early research on modeling such data appeared mainly in the econometrics literature. Tobin (1958) proposed a censored regression model to describe household expenditures on durable goods. This model is now commonly referred to as the Tobit model.[1] The term "Tobit" arose from its similarities in derivation to the probit model, based on a normal latent variable construction described below. Since then, related literature contains numerous econometric applications as well as various generalizations of the Tobit model (e.g., Cragg 1971, Amemiya 1973, Gronau 1974, Heckman 1974, 1979). These all posit an underlying normal random variable that is censored by a random mechanism.

An alternative strand of literature for semicontinuous data does not assume an underlying normal distribution. Duan et al. (1983) proposed a two-part model to fit data on expenditures for medical care. Jørgensen (1987) proposed a compound Poisson exponential dispersion model for semicontinuous data. Saei, Ward, and McGilchrist (1996) applied an ordinal response model that requires grouping the response outcomes into categories. The Tobit model and these alternative models are described in the following subsections.

1.1.1 Tobit Models

For response variable Y, let y_i denote the observation for subject i, $i = 1, \ldots, n$. The Tobit model assumes an underlying normally distributed variable Y_i^*

[1] James Tobin, Sterling Professor Emeritus of Economics at Yale University, won the 1981 Nobel Prize in Economics; he died on March 11, 2002.

such that

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0. \end{cases}$$

When $y_i^* \leq 0$, its value is unobserved.

Including explanatory variables, the model assumes that the underlying variable is generated by

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \tag{1.1}$$

where $\boldsymbol{x}_i$ is a column vector of explanatory variable values for subject i and $\{u_i\}$ are independent from a normal $N(0, \sigma^2)$ distribution. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function (*cdf*) and the probability density function (*pdf*) of the $N(0,1)$ distribution. For the Tobit model, the probability of a zero response is

$$P(Y_i = 0) = P(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i \leq 0) = P(u_i \leq -\boldsymbol{x}_i'\boldsymbol{\beta}) = \Phi\left(\frac{-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) = 1 - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right).$$

Conditional on $y_i > 0$, its probability density function is

$$f(y_i; \boldsymbol{\beta}, \sigma) = \sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right).$$

Thus, the likelihood function for a sample of n independent observations is

$$L(\boldsymbol{\beta}, \sigma) = \left[\prod_{y_i=0}\left\{1 - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\}\right]\left[\prod_{y_i>0}\sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]. \tag{1.2}$$

Tobin (1958) used a Newton-Raphson algorithm to find the maximum likelihood (ML) estimates of $\boldsymbol{\beta}$ and σ. Amemiya (1984) presented a comprehensive survey of the Tobit model and its generalizations.

The Tobit model assumes normality for the distribution of the error term, with constant variance. In many applications this is unrealistic. When the model form is correct but the distribution of u_i is not normal, the ML estimators are inconsistent (Robinson 1982).

Powell (1986) proposed semi-parametric estimation for the Tobit model. He used a symmetrically trimmed least squares (STLS) estimator. This assumes that $\{u_i\}$ are symmetrically distributed about zero. The STLS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{STLS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} I(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)[\min(y_i, 2\boldsymbol{x}_i'\boldsymbol{\beta}) - \boldsymbol{x}_i'\boldsymbol{\beta}]^2,$$

where I is the indicator function. For a given $\boldsymbol{\beta}$, the sum in this expression deletes the observations with $\boldsymbol{x}_i'\boldsymbol{\beta} \leq 0$. When $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$, the lower tail of the distribution of Y_i is censored at zero; symmetrically censoring the upper tail of the distribution (essentially by replacing y_i by $\min\{y_i, 2\boldsymbol{x}_i'\boldsymbol{\beta}\}$) restores the symmetry of distribution of Y^*. The resulting estimator $\hat{\boldsymbol{\beta}}_{STLS}$ is consistent and asymptotically normal under the symmetrical distribution assumption (Powell 1986). An iterative procedure yields $\hat{\boldsymbol{\beta}}_{STLS}$.

Yoo, Kim, and Lee (2001) used this method with the bootstrap to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{STLS}$. For M bootstrap replications with estimate $\hat{\boldsymbol{\beta}}_j$ in replication j, their estimate is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{M} \sum_{j=1}^{M} (\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_{STLS})(\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_{STLS})',$$

where $\bar{\boldsymbol{\beta}}_{STLS} = (1/M)\sum_{j=1}^{M} \hat{\boldsymbol{\beta}}_j$. In an empirical study, Yoo et al. showed that semi-parametric estimation significantly outperforms estimation assuming normality (i.e., the Tobit model).

1.1.2 Two-Part Models

The Tobit model allows the same underlying stochastic process to determine whether the response is zero or positive as well as the value of a positive response. That is, the same parameters influence whether the outcome is zero or positive as well as the magnitude of the outcome, conditional on its being positive. The next two subsections discuss "two-part models" that allow the two components to have different parameters.

Without assuming an underlying normal distribution, Duan et al. (1983) proposed a two-part model that uses two equations to separate the modeling into two stages. The first stage refers to whether the response outcome is positive. Conditional on its being positive, the second stage refers to its level.

The first part is a binary model for the dichotomous event of having zero or positive values, such as the logistic regression model

$$\operatorname{logit}[P(Y_i = 0)] = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1. \tag{1.3}$$

Conditional on a positive value, the second part assumes a log-normal distribution; that is,

$$\log(y_i | y_i > 0) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 + \epsilon_i, \tag{1.4}$$

where ϵ_i is distributed as $N(0, \sigma^2)$. The likelihood function for this two-part model is

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) &= \Big[\prod_{y_i=0} P(y_i = 0)\Big]\Big[\prod_{y_i>0} P(y_i > 0) f(y_i | y_i > 0)\Big] \\ &= \Big[\prod_{y_i=0} \frac{e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}\Big]\Big[\prod_{y_i>0} \frac{1}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}} \sigma^{-1} \phi\Big(\frac{\log(y_i) - \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}{\sigma}\Big)\Big]. \end{aligned}$$

Duan et al. (1983) showed that the likelihood function has a unique global maximum. ML calculations are relatively simple, because the likelihood function factors into two terms. The first term has only the logit model parameters,

$$L_1(\boldsymbol{\beta}_1) = \Big[\prod_{y_i=0} e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}\Big]\Big[\prod_{i=1}^{n} \frac{1}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}\Big].$$

The second term involves only the parameters of the second model part,

$$L_2(\boldsymbol{\beta}_2, \sigma) = \prod_{y_i>0} \sigma^{-1} \phi\Big(\frac{\log(y_i) - \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}{\sigma}\Big).$$

One can obtain ML estimates by separately maximizing the two terms. Duan et al. (1983) applied this model to describe demand for medical care. For another application, see Grytten, Holst, and Laake (1993).

1.1.3 Sample Selection Models

Heckman (1974, 1979) extended the Tobit model to a two-part model. His model has been commonly applied to model sample selection and the related potential bias. There are many variants of sample selection models. We use the version by van de Ven and van Praag (1981) to illustrate. For observation i, let $\{(u_{1i}, u_{2i})\}$ be *i.i.d.* from a bivariate $N(\mathbf{0}, \mathbf{\Sigma})$ distribution, where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The model assumes that

$$I_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + u_{1i}, \tag{1.5}$$

$$y_i^* = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + u_{2i}, \tag{1.6}$$

$$\begin{aligned} y_i &= \exp(y_i^*) \quad \text{if } I_i > 0, \\ &= 0 \quad \text{if } I_i \leq 0. \end{aligned}$$

When $I_i > 0$, $y_i > 0$ is observed and $y_i^* = \log(y_i)$; when $I_i \leq 0$, $y_i = 0$ is observed and y_i^* is "missing". The covariate and parameter vectors $(\boldsymbol{x}_{1i}, \boldsymbol{\beta}_1)$ for I_i may differ from $(\boldsymbol{x}_{2i}, \boldsymbol{\beta}_2)$ for y_i^*. Two estimation methods employed with this model are ML and a two-step procedure due to Heckman (1979).

For ML estimation, the likelihood function of the model is given by

$$\begin{aligned}
L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \Sigma) &= \Big[\prod_{y_i=0} P(I_i \le 0)\Big]\Big[\prod_{y_i>0} f(y_i^*|I_i > 0)P(I_i > 0)\Big] \\
&= \Big[\prod_{y_i=0} P(I_i \le 0)\Big]\Big[\prod_{y_i>0} \int_0^{\infty} f(y_i^*, I_i)dI_i\Big] \\
&= \Big[\prod_{y_i=0}\{1 - \Phi\Big(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\Big)\}\Big] \\
&\quad \times\Big[\prod_{y_i>0} \Phi\{\Big(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1} + \frac{\log(y_i) - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_{12}^{-1}\sigma_1\sigma_2^2}\Big) \\
&\quad \times(1 - \sigma_{12}^2\sigma_1^{-2}\sigma_2^{-2})^{-\frac{1}{2}}\}\sigma_2^{-1}\phi\Big(\frac{\log(y_i) - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\Big)\Big].
\end{aligned}$$

An iterative method can be used to find the ML estimates.

Heckman's two-step procedure does not perform as well as the ML estimators. But this method is very simple and easy to implement. It is widely used and has become the standard estimation procedure for empirical microeconometrics studies. With the two-step procedure, the subsample regression function for Y_i^* is

$$E[Y_i^*|\boldsymbol{x}_{2i}, I_i > 0] = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + E[u_{2i}|u_{1i} > -\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1] = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_i, \qquad (1.7)$$

where $\lambda_i = \phi(z_i)/\Phi(z_i)$, and $z_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1/\sigma_1$. So, we have

$$\begin{aligned}
\log(Y_i) &= E[Y_i^*|\boldsymbol{x}_{2i}, I_i > 0] + \epsilon_i \\
&= \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_i + \epsilon_i,
\end{aligned}$$

where Heckman (1979) showed that ϵ_i has mean 0 and variance

$$\sigma_2^2[(1 - \rho^2) + \rho^2(1 + z_i\lambda_i - \lambda_i^2)],$$

where $\rho^2 = \sigma_{12}^2/(\sigma_1^2\sigma_2^2)$. One can estimate the parameters $\boldsymbol{\beta}_1$ and σ_1 by a probit model using the full sample. Therefore, z_i and hence λ_i can be easily estimated. The estimated value of λ_i is used as a regressor in equation (1.7). Then one can estimate $\boldsymbol{\beta}_2$ using least squares.

Duan et al. (1983, 1984) pointed out that the model has poor numerical and statistical properties. The likelihood function may have non-unique local maxima (Olsen 1975), and computations are more involved than in the Duan et al. (1983) two-part model. The model relies on untestable assumptions in that the censored data are unobservable, so standard diagnostic methods based on the empirical error distribution cannot be applied. When a high correlation exists between λ and $\boldsymbol{x}_2$, the estimator in the sample selection model is very nonrobust. Some researchers have suggested that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ should not have variables in common, but this is not realistic in practice.

Both the Duan et al. (1983) two-part model and Heckman's sample selection model use two equations to separately model whether the outcome is positive and the magnitude of a positive response. The sample selection model posits an underlying bivariate normal error. It estimates an unconditional equation that describes the level that subjects would have if they all had outcomes. The two-part model estimates a conditional equation that describes only the level of outcomes for those that truly are positive. The econometrics literature contains discussion comparing the sample selection model and the two-part model. See, for instance, Duan et al. (1983, 1984), Manning et al. (1987), and Leung and Yu (1996).

1.1.4 Compound Poisson Exponential Dispersion Models

Jørgensen (1987, 1997) proposed using a single distribution from the exponential dispersion family to analyze semicontinuous data. This distribution is a type of compound Poisson distribution. The exponential dispersion family, which is used in generalized linear models, has form

$$f(y_i;\, \theta_i, \phi) = c(y_i, \phi) \exp\left(\frac{\theta_i y_i - b(\theta_i)}{\phi}\right). \tag{1.8}$$

It is characterized by its variance function $V(\mu_i)$, expressed in terms of the mean μ_i (Jørgensen 1987). For this family, θ relates to μ by $\mu = \partial b(\theta)/\partial\theta$. An important class of exponential dispersion models uses the power function, $V(\mu) = \mu^p$. When $p = 1$, this is the Poisson distribution.

Jørgensen (1997) applied this family for $1 < p < 2$, for which

$$b(\theta_i) = (\frac{\alpha-1}{\alpha})(\frac{\theta_i}{\alpha-1})^{\alpha},$$

where $\alpha = (p-2)/(p-1)$, and

$$c(y_i, \phi) = \begin{cases} \frac{1}{y_i}\sum_{n=1}^{\infty} \frac{b^n(-\phi/y_i)}{\phi^n \Gamma(-\alpha n) n!} & y_i > 0 \\ 1 & y_i = 0. \end{cases}$$

For this distribution,

$$\mu_i = \partial b(\theta_i)/\partial\theta_i = \left(\frac{\theta_i}{\alpha-1}\right)^{\alpha-1}.$$

Jørgensen (1997) showed that when $1 < p < 2$, this distribution results from the compound Poisson construction,

$$Y_i = \sum_{j=0}^{N_i} W_{ij}, \tag{1.9}$$

where N_i has a Poisson$(b(\theta_i)/\phi)$ distribution and W_{ij} has a gamma$(\alpha\phi/\theta_i, -\alpha)$ distribution. When N_i and $\{W_{ij}\}$ are independent, $P(Y_i = 0) = P(N_i = 0)$. Given $N_i > 0$, the distribution of Y_i is continuous on the positive real line.

With link function $g()$, one can specify a model for the mean response as $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. Obtaining the ML estimator for $\boldsymbol{\beta}$ does not involve $c(y_i, \phi)$. When p is known, this model can be fitted with software for generalized linear models. Normally, however, p would itself be unknown and need to be estimated. Since it occurs (through α) in the infinite sum and gamma function in $c(y_i, \phi)$, estimating it can be computationally difficult (Jørgensen 1987). Alternative moment-based estimation may perform well. Tweedie (1984) suggested an estimate of p based on a single random sample as $\hat{p} = \hat{k}_1\hat{k}_3\hat{k}_2^{-2}$, where $\hat{k}_t$ is an estimate of cumulant t of the distribution. Jørgensen proposed a possible generalization of this approach for a regression model. Let $\boldsymbol{y}$ and $\hat{\boldsymbol{\mu}}$ represent vectors of observations and fitted values. A moment estimator for ϕ is $\hat{\phi} = \boldsymbol{X}^2/(n-k)$, where k is the number of unknown parameters and $\boldsymbol{X}^2 = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T V(\hat{\boldsymbol{\mu}})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

1.1.5 Ordinal Threshold Models

Saei, Ward, and McGilchrist (1996) suggested grouping the possible outcome values into K ordered categories and applying an ordinal response model. Let Y_g be the grouped response variable. The threshold model for an ordinal response posits an unobservable variable Z, such that one observes $Y_g = k$ (i.e., in category k) if Z is between θ_{k-1} and θ_k. Suppose that Z has a cumulative distribution function $G(z - \eta)$, where η is related to explanatory variables by

$$\eta = \boldsymbol{x}'\boldsymbol{\beta}.$$

Then,

$$P(Y_g \leq k) = P(Z \leq \theta_k) = G(\theta_k - \boldsymbol{x}'\boldsymbol{\beta}).$$

The threshold model then follows, by which

$$G^{-1}[P(Y_g \leq k; \boldsymbol{x})] = \theta_k - \boldsymbol{x}'\boldsymbol{\beta}, \quad k = 1, 2, \ldots, K-1. \tag{1.10}$$

That is, the inverse of the *cdf* serves as the link function.

In application with semicontinuous data and a clump at 0, one would take the first category to be the 0 outcome, and then one would select cutpoints on the positive outcome scale to define the other $K - 1$ categories. Assuming that G is logistic leads to a logit model for the cumulative probabilities, called a cumulative logit model. Assuming that G is normal leads to a cumulative probit model (McCullagh 1980). A score test is available to check the assumption that covariate effects are the same for each cutpoint (Peterson and Harrell 1990). Chang and Pocock (2000) applied the cumulative logit model for modeling the amount of personal care for the elderly.

This model has the simplicity of a single model to handle the clump at 0 and the positive outcomes. Elements of $\boldsymbol{\beta}$ summarize effects overall, rather than conditional on the response being positive. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for two-part models one needs to average results from the two components of the model to make an unconditional comparison (e.g., to estimate $E(Y)$ for the groups). Two obvious concerns with this model are that the way the positive scale is collapsed into categories is arbitrary, and by grouping the data one loses some information.

1.1.6 Advantages and Disadvantages of Existing Approaches

The Tobit model was the first to deal with semicontinuous data. The sample selection model extends the Tobit model to allow different coefficients to affect the two components. Both models assume an underlying normal random variable that is censored by a random mechanism. These models are sometimes suitable for modeling a limited or censored response variable. When zeros represent actual outcome values instead of censored or missing values, the underlying normal assumption becomes dubious. By contrast, the Duan et al. (1983) two-part model has several appealing properties, including a well-behaved likelihood function and more appropriate interpretations than the Tobit and Heckman models if the zeros are true values.

The compound Poisson exponential dispersion model makes it possible to analyze data with a single model that includes both aspects described in the two-part model. In this sense, it is relatively simple. Given the power p in the variance function, this model is easy to fit, but otherwise the model seems problematic. It does not seem to have received attention in practice other than in Jørgensen's work. Ordinal response models also can model the zero and non-zero values in one model, and they are simple to fit. A drawback is that they model grouped data instead of the original data.

Of these models, it seems to us that the Duan et al. (1983) two-part model is a reasonable choice for many applications. Compared with other models we have discussed, this model addresses the data in their original form, is simple to fit, and is relatively simple to interpret.

1.2 Models for Zero-Inflated Count Data

Count responses with a relatively large clump at zero can occur in many situations (e.g., Cameron and Trivedi 1998, pp. 10-15). Having a large number of observations at zero is not by itself sufficient to rule out a particular discrete distribution. However, often the remaining counts show considerable variability, which is inconsistent with the Poisson distribution (for which the mean determines both the variance and the probability at 0). This may be caused by overdispersion due to unobserved heterogeneity. Then, a distribution that allows the Poisson mean to vary at fixed values of predictors may be appropriate. Examples are the negative binomial regression model (which can be derived with a gamma mixture of Poisson means) and the generalized linear mixed model that adds a normal random effect to a model for the log of the Poisson mean. See, for instance, Cameron and Trivedi (1998) and Chapter 13 of Agresti (2002) for discussion of such approaches.

Sometimes such simple models for overdispersion are themselves inadequate. For instance, the data might be bimodal, with a clump at zero and a separate hump around some considerably higher value. This might happen for variables for which a certain fraction of the population necessarily has a zero outcome, and the remaining fraction follows some distribution having positive probability of a zero outcome. This happens for variables referring to the number of times one takes part in a certain activity, when some subjects never do so and others may occasionally not do so. Examples are the number of papers one published in the previous year (for a sample of professors), and the number of times one exercised in a gym in the previous month. For such zero-clumped data, standard discrete distributions are suspect. The above representation of two types of subjects leads naturally to a mixture model, some examples of which are presented in this section on the modeling of zero-inflated count data.

1.2.1 Zero-Inflated Discrete Distributions

Lambert (1992) introduced zero-inflated Poisson (ZIP) regression models to account for overdispersion in the form of excess zero counts for the Poisson distribution. Since her article, zero-inflated discrete models have been developed and applied in the econometrics and statistics literature.

Lambert's model treats the data as a mixture of zeros and outcomes of Poisson variables. For subject i, she assumed that

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - p_i. \end{cases}$$

The resulting distribution has

$$\begin{aligned} P(Y_i = 0) &= p_i + (1 - p_i)e^{-\lambda_i}, \\ P(Y_i = j) &= (1 - p_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!}, \qquad j = 1, 2, \ldots . \end{aligned}$$

With explanatory variables, the parameters are themselves modeled by

$$\text{logit}(p_i) = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2. \tag{1.11}$$

The log likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= \sum_{y_i=0} \log[e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1} + \exp(-e^{\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2})] + \sum_{y_i>0}(y_i\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 - e^{\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}) \\ &\quad - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1}) - \sum_{y_i>0} \log(y_i!). \end{aligned}$$

A latent class construction that yields this model posits an unobserved binary variable Z_i. When $Z_i = 1$, $y_i = 0$, and when $Z_i = 0$, Y_i is Poisson(λ_i). Lambert (1992) suggested using the EM algorithm for ML estimation of the parameters, treating z_i as a missing value.

Hall (2000) adapted Lambert's method to an upper-bounded count setting to yield a zero-inflated binomial model. With upper bound for Y_i of n_i, he took

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{binomial}(n_i, \pi_i) & \text{with probability } 1 - p_i. \end{cases}$$

He modeled p_i with $\text{logit}(p_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1$ and modeled π_i with $\text{logit}(\pi_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$, using the EM algorithm to obtain ML estimates.

In practice, overdispersion is common with count data, even conditional on a positive count or for a component of a latent class model. The equality of mean and variance assumed by the ZIP model, conditional on $Z_i = 0$, is often not realistic. Zero-inflated negative binomial models would likely often be more appropriate that ZIP models. Grogger and Carson (1991) used zero-truncated Poisson models to fit data simulated from zero-truncated negative binomial distributions. They observed biases of estimated parameters up to 30 percent. Similar arguments extend to zero-inflated models. With an inappropriate Poisson assumption, standard error estimates can be biased very dramatically. Ridout, Hinde and Demetrio (2001) provided a score test for testing zero-inflated Poisson models against the zero-inflated negative binomial alternative. For an application of the zero-inflated negative binomial model, see Shankar, Milton, and Mannering (1997).

With more than a single unusually high probability, extensions of zero-inflated count models may be needed. For instance, in studying Swedish female fertility, Melkersson and Rooth (2000) inspected the number of births for a sample of women. They found more 0 and 2 outcomes than expected in a standard count data model. They used a multinomial logit model to estimate the extra probabilities of zero and two children.

1.2.2 Hurdle Models

The hurdle model is a two-part model for count data proposed by Mullahy (1986). One part of the model is a binary model, such as logistic or probit regression, for whether the response outcome is zero or positive. If the outcome is positive, the "hurdle is crossed." Conditioning on a positive outcome, to analyze its level the second part uses a truncated model that modifies an ordinary distribution by conditioning on a positive outcome. This might be a truncated Poisson or truncated negative binomial. Applications of such models have been given by Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997), and Gurmu and Trivedi (1996).

Suppose we use a logistic regression for the binary process and a truncated Poisson model for the positive outcome; that is,

$$\text{logit}[P(Y_i = 0)] = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2. \tag{1.12}$$

The log likelihood then has two components:

$$\begin{aligned} \ell_1(\boldsymbol{\beta}) &= \sum_{y_i=0}[\log P_1(y_i = 0; \boldsymbol{\beta}_1, \boldsymbol{x}_{1i})] + \sum_{y_i>0}[\log(1 - P_1(y_i = 0; \boldsymbol{\beta}_1, \boldsymbol{x}_{1i}))] \\ &= \sum_{y_i=0} \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}) \end{aligned}$$

is the log-likelihood function for the binary process, and

$$\ell_2(\boldsymbol{\beta}_2) = \sum_{y_i>0}[y_i\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 - e^{\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2} - \log(1 - e^{-e^{\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}})] - \sum_{y_i>0} \log(y_i!)$$

is the log-likelihood function for the truncated model. The joint log-likelihood function is

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \ell_1(\boldsymbol{\beta}_1) + \ell_2(\boldsymbol{\beta}_2).$$

One can maximize this by separately maximizing ℓ_1 and ℓ_2.

In some applications, the data may have a long right tail reflecting some extremely large positive counts. Gurmu (1997) proposed a semi-parametric hurdle model for a highly skewed distribution of counts. It is based on a Laguerre series expansion for the unknown density of the unobserved heterogeneity.

1.2.3 Finite Mixture Models

Another approach for zero-inflated count data uses a finite mixture model. It assumes that the response comes from a mixture of latent distributions. With C latent groups, the mixture density is

$$f(y_i; \boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c f_c(y_i; \theta_c), \quad y_i = 0, 1, 2, \ldots, \tag{1.13}$$

where π_c is the true proportion in group c, $f_c(y_i; \theta_c)$ is the mass function (e.g., Poisson or negative binomial) for group c, and $\{\pi_c\}$ and $\{\theta_c\}$ are unknown parameters. The zero-inflated count models of Sec. 1.2.1 are special cases of the finite mixture model in which one of the mixture mass functions is degenerate at zero. The more general mixture model allows for additional population heterogeneity but avoids the sharp dichotomy between the population of zeros and non-zero counts.

One approach to fitting a finite mixture model relates it to latent class analysis (Aitkin and Rubin 1985). Let d_{ic} denote an indicator to represent whether y_i comes from latent group c, with $\sum_c d_{ic} = 1$. Assume that $\{(y_i, d_{i1}, \ldots, d_{iC}),\ i = 1, \ldots, n\}$ are independent, such that $\{d_{ic}, c = 1, \ldots, C\}$ have the multinomial distribution

$$\prod_{c=1}^{C} \pi_c^{d_{ic}},$$

and conditional on their values, y_i has probability mass function

$$\sum_{c=1}^{C} d_{ic} f(y_i;\theta_c) = \prod_{c=1}^{C} f(y_i;\theta_c)^{d_{ic}}, \quad y_i = 0, 1, 2, \ldots.$$

Then, the likelihood function is

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\sum_{c=1}^{C} \pi_c^{d_{ic}} f(y_i;\theta_c)^{d_{ic}} \right].$$

Treating $\{d_{ic}\}$ as missing data, one can use the EM algorithm to fit the model.

Deb and Trivedi (1997) used a finite mixture model to study the demand for medical care by the elderly. They found that a two-point mixture negative binomial model fits better than the standard negative binomial model and its hurdle extension. Wedel et al. (1993) applied a finite Poisson mixture model to analyze the effects of direct marketing on book selling. Gerdtham and Trivedi (2001) used a finite Poisson mixture model in studying the equity issue in Swedish health care.

1.2.4 Neyman Type A Distribution

Dobbie and Welsh (2001b) proposed modeling zero-inflated count data using the Neyman type A distribution. This distribution is a compound Poisson-Poisson mixture. For observation i, let N_i denote a Poisson variate with expected value λ_i. Conditional on N_i, let W_{it} $(t = 1, \ldots, N_i)$ denote independent observations from a Poisson distribution with expected value ϕ_i. The model expresses Y_i using the decomposition,

$$Y_i = \sum_{t=0}^{N_i} W_{it}, \quad i = 1, 2, \ldots, n.$$

The probability mass function for Y_i is

$$\begin{aligned} P(Y_i = y_i) &= \sum_{j=0}^{\infty} \left[P(\sum_{t=0}^{N_i} W_{it} = y_i | N_i = j) P(N_i = j) \right] \\ &= \sum_{j=0}^{\infty} \left[\frac{e^{-j\phi_i}(j\phi_i)^{y_i}}{y_i!} \right] \left[\frac{e^{-\lambda_i}\lambda_i^j}{j!} \right] \\ &= \frac{e^{-\lambda_i}\phi_i^{y_i}}{y_i!} \sum_{j=0}^{\infty} \frac{(\lambda_i e^{-\phi_i})^j j^{y_i}}{j!}. \end{aligned}$$

Using this distribution, one can form a model that relates λ_i and ϕ_i to explanatory variables through

$$\log(\lambda_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1,$$

$$\log(\phi_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2.$$

Since $E(Y_i) = \lambda_i\phi_i$,

$$\log[E(Y_i)] = \log(\lambda_i) + \log(\phi_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2.$$

Dobbie and Welsh used a four-step procedure with the Newton-Raphson algorithm to estimate parameters, iterating between estimating $\boldsymbol{\beta}_1$ for a given $\boldsymbol{\beta}_2$ and estimating $\boldsymbol{\beta}_2$ for a given $\boldsymbol{\beta}_1$. The infinite sums in the density function make model-fitting complicated. They applied it to model the abundance of Leadeater's Possum in mountain ash forests of southeastern Australia. Here, λ_i denotes the mean number of possum clusters per site, and ϕ_i denotes the average number of possums per cluster.

1.2.5 Advantages and Disadvantages of Existing Approaches

The zero-inflated model and the hurdle model are similar. The zero-inflated models are more natural when it is reasonable to think of the population as a mixture, with one set of subjects that necessarily has a 0 response. However, they are more complex to fit, as the model components must be fitted simultaneously.

By contrast, one can separately fit the two components in the hurdle model. The hurdle model is also suitable for modeling data with *fewer* zeros than would be expected under standard distributional assumptions.

The finite mixture model is semi-parametric. If the observations can realistically be viewed as being drawn from different populations, this approach is attractive. A potential disadvantage with this model is that it may overestimate the number of components when there is a lack of model fit. The Neyman type A model makes it possible to fit the data using a single distribution. However, it is not a member of the exponential family, so the mathematical and inferential advantages associated with this family are not available, and model fitting is complicated by the infinite sum in the mass function.

1.3 Methods for Compliance Data Analysis

In compliance data analysis, the compliance response is treated as a continuous proportion, except for a discrete mass at 0 and 1. Two characteristics of compliance data make them hard to analyze. The first is that the response variable is a continuous proportion distributed on the unit interval. The second is that compliance data typically have two clumps, at 0 and at 1. The logit transformation can be used to transform the response values between zero and one to the real line, as is often done with compositional data (Aitchison and Shen 1980) and continuous proportion data (Bartlett 1937). However, this transformation can not handle 0% compliances and 100% compliances.

Most studies on compliance data appear in medical journals. In comparing patient compliance of several groups without considering other covariates, most of the medical studies (e.g., Melikian et al. 2002) have used the one-way analysis of variance (ANOVA) method. This method is based on the assumption that the data are sampled from normal distributions and the variances of all groups are equal. The one-way ANOVA method is robust to deviations from the assumptions for large sample sizes or when there are no outliers and the distribution is roughly symmetric. However, compliance data usually contain a substantial proportion of 0% and 100% compliances, and the continuous proportions are usually skewed and do not have a bell-shaped distribution. For small sample sizes, the central limit theorem is not applicable. For these reasons, the P-value from the one-way ANOVA model may be inaccurate. Moreover, with ANOVA we cannot estimate the effects of covariates (such as socio-demographic and behavioral factors) on the compliances.

For pairwise comparison, most papers in medical journals have applied t tests with Bonferroni's procedure (e.g., Waterhouse et al. 1993). The t test is also based on a normal distribution assumption. Like ANOVA, the P-values obtained from these tests are not reliable. A few medical papers use nonparametric methods in comparing compliance data from several groups. Detry et al. (1995) used the Wilcoxon rank sum test in comparing the median of compliance data from two groups. This is a nonparametric analogue to the two-sample t test. It does not make normal assumptions about the population distributions. Sherman et al. (2000) used a nonparametric bootstrap percentile confidence intervals method to construct 95% confidence intervals for mean compliance rates. Differences in compliance rates were considered significant when the 95% confidence intervals do not overlap. This method is not appropriate, since the standard error of the difference does not equal the sum of the standard errors of the two groups.

When considering the relationship between patient compliance and covariates, the most frequently used methods are analysis of covariance (ANCOVA) or two-way ANOVA. Since these two methods are based on the same assumptions as the one-way ANOVA, they have the same problem as the one-way ANOVA method. A few papers (e.g., Waterhouse et al. 1993) classified compliance as a dichotomous variable – adherence ($>$ 80%) and non-adherence ($\leq$ 80%), or low-compliance

($\leq$ 50%) and high-compliance ($>$ 50%). Then they used logistic regression to fit the classified data. Dichotomization of compliance is not always possible in clinical trials, as thresholds of compliance for acceptable therapeutic effects are not known in advance. Therefore, treating compliance as a continuous variable or an ordinal variable having more than two categories may be more appropriate.

In contrast with the many models for data with clumping at zero, little attention has been paid to specialized appropriate models for compliance data.

1.4 Modeling Repeated Measurements of Zero-Clumped Data

Compared with the substantial literature on cross-sectional observations of data with clumping at zero, few papers have discussed the modeling of clustered, correlated observations, such as occur with longitudinal data. This section surveys this literature.

1.4.1 Repeated Measurements of Semicontinuous Data

Cowles, Carlin, and Connett (1996) and Hajivassiliou (1994) extended the Tobit model and the sample selection model to longitudinal data. Both models assume an underlying normal distribution, which is dubious in most applications, especially when zeros represent actual responses instead of censored or missing values. We do not discuss their approaches here. Olsen and Schafer (2001) extended the two-part model of Duan et al. (1983) to longitudinal data. We describe their model next.

Let y_{ij} be the semicontinuous response for subject (or cluster) i ($i = 1, \ldots, n$) at occasion j ($j = 1, \ldots, t_i$). The first part of the model is a logistic random effects model for the dichotomous event of having zero or positive values. Suppose that

$$y_{ij} \begin{cases} = 0 & \text{with probability } p_{ij} \\ \neq 0 & \text{with probability } 1 - p_{ij}. \end{cases}$$

Let $\{\boldsymbol{b}_{1i}\}$ be random effects to account for within-subject correlation. Conditional on $\boldsymbol{b}_{1i}$, we assume that

$$\operatorname{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i}, \tag{1.14}$$

where $\boldsymbol{x}_{1ij}$ and $\boldsymbol{z}_{1ij}$ are covariate vectors pertaining to the fixed effects $\boldsymbol{\beta}_1$ and random effects $\boldsymbol{b}_{1i}$. In practice, the simple random intercept form of model is often adequate, in which $\boldsymbol{b}_{1i} = b_{1i}$ is univariate and $\boldsymbol{z}_{1ij} = 1$.

In the second part of the model, let

$$y^*_{ij} = \begin{cases} y_{ij}, & \text{if } y_{ij} > 0 \\ \text{unspecified}, & \text{if } y_{ij} = 0\ . \end{cases}$$

When y^*_{ij} is positive, conditional on a random effect $\boldsymbol{b}_{2i}$ the model assumes that Y^*_{ij} follows a log-normal distribution. Thus, the model for the positive outcomes is

$$\log(y^*_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i} + \epsilon_{ij}, \tag{1.15}$$

where the residuals $\{\epsilon_{ij}\}$ are assumed to be independent from $N(0, \sigma^2)$. Again, often the simple random intercept form of model is often adequate, in which $\boldsymbol{b}_{2i} = b_{2i}$ is univariate and $\boldsymbol{z}_{2ij} = 1$.

When the response is observed at repeated times, a high level of a positive response at one time may affect the probability of a positive outcome at another time. So, one can tie the two parts of the model together by taking the random effects from the two parts as jointly normal and correlated,

$$\boldsymbol{b}_\imath = \begin{pmatrix} \boldsymbol{b}_{1i} \\ \boldsymbol{b}_{2i} \end{pmatrix} \sim N\left(\boldsymbol{0},\ \boldsymbol{\Sigma}\right),$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{b_1b_1} & \boldsymbol{\Sigma}_{b_1b_2} \\ \boldsymbol{\Sigma}_{b_2b_1} & \boldsymbol{\Sigma}_{b_2b_2} \end{pmatrix}.$$

To fit the model, one first obtains a marginal likelihood by integrating out the random effects. However, these integrals are analytically intractable, so the marginal likelihood does not have a closed-form expression. Numerical or stochastic approximation of the integrals is needed, as in the fitting of generalized linear mixed models (e.g., Agresti 2002, Chapter 12). With univariate random intercepts, numerical approximation using Gauss-Hermite quadrature, which approximates the integral by a finite sum, should be adequate. Then one can maximize the approximated likelihood using standard optimization methods such as Newton–Raphson.

Olsen and Schafer (2001) studied many fitting methods. They compared Markov chain Monte Carlo (MCMC), the EM algorithm, penalized quasi-likelihood (PQL), Gauss-Hermite quadrature, and Laplace approximations. Simulations by Raudenbush et al. (2000) showed that a high-order Laplace approximation can be as accurate as Gauss-Hermite quadrature yet is much faster than the other methods. Olsen and Schafer noted that it took MCMC and EM algorithms more than one day to obtain accurate estimates for their example, while the sixth-order Laplace method needed less than one minute.

Saei et al. (1996) extended the ordinal threshold model to analyze clustered semicontinuous data. Again, this requires breaking the continuous scale into categories. Let $y_{ij,g}$ be the grouped response for observation j on subject i. Let G be the cumulative distribution function for an underlying unobservable variable. For outcome category k, the model assumes

$$P(Y_{ij,g} \leq k) = G(\theta_k - \eta_{ij}),$$

where

$$\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i$$

for a vector $\boldsymbol{b}_i \sim N(\mathbf{0}, \Sigma)$ of random effects that account for within-subject correlation. They took G to be the standard normal *cdf*, yielding a cumulative probit model, and they used penalized quasi-likelihood (PQL) to fit the model.

1.4.2 Repeated Measurements of Zero-Inflated Data

As with semicontinuous data, there is little literature on modeling clustered zero-inflated count data. Hall (2000) extended the zero-inflated Poisson and zero-inflated binomial models to handle longitudinal data, adding random effects to account for the within-subject dependence.

Hall assumed that $Y_{ij} = 0$ with probability p_{ij} and $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ with probability $1 - p_{ij}$. The parameters p_{ij} and λ_{ij} are modeled by

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1, \tag{1.16}$$

$$\log(\lambda_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + b_i, \tag{1.17}$$

where $b_i \sim N(0, \sigma^2)$ is a random effect. The Poisson(λ_{ij}) distribution applies conditional on b_i; unconditionally, there is overdispersion relative to the Poisson when $\sigma > 0$. The log-likelihood function for the longitudinal zero-inflated Poisson model is

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) = \sum_{i=1}^{n} \left[\log \int_{-\infty}^{+\infty} [\prod_{j=1}^{t_i} P(Y_{ij} = y_{ij}|b_i)]\phi(b_i)db_i \right].$$

Hall employed the EM algorithm with Gauss-Hermite quadrature to fit the model. In a corresponding longitudinal zero-inflated binomial model, Hall assumed that Y_{ij} is binomial(n_i, π_{ij}) with probability $1 - p_{ij}$ (conditional on a random effect b_i), where $\text{logit}(\pi_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + b_i$.

Note that Hall's model does not have a random effect for the part of the model determining the zero inflation. By contrast, Yau and Lee (2001) proposed adding a pair of uncorrelated normal random effects (b_{1i}, b_{2i}) for the two components of a hurdle model. They used a logistic model for the probability p_{ij} of a positive outcome. Conditional on positive outcome, they applied a log linear model for the mean λ_{ij} in a truncated (conditionally) Poisson distribution. That is,

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + b_{1i}, \tag{1.18}$$

$$\log(\lambda_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + b_{2i}, \tag{1.19}$$

With uncorrelated random effects, the two components of the hurdle model can be fitted separately. Yau and Lee used a penalized quasi-likelihood approach for this.

1.5 Outline of Dissertation

In this dissertation, we plan to develop methods for modeling a few special cases of data with clumps, which include repeated zero-inflated count data, cross-sectional compliance data, and repeated compliance data. In Chapter 2 we first use two simulation studies to address the advantage of using the hurdle model over the zero-inflated count model. Then we present two types of random effects models for repeated measures on zero-inflated count data. The first model is a correlated random effects hurdle model, which separately handles the zero observations and the positive counts. In maximum likelihood model fitting, we consider both a normal distribution and a nonparametric approach for the random effects. We also discuss a special type of the hurdle model, which can be used to test the existence of zero-inflation. The second model is a cumulative logit model with random effects, which has the simplicity of using a single model to handle the zero-inflation problem. We illustrate the proposed methods with an example from an occupational injury prevention program. We also consider the identifiability issue of the ZIP model in this chapter.

Chapter 3 concentrates on modeling cross-sectional compliance data. We introduce a two-part model to analyze compliance data. We then extend the two-part model to a more general model – the mixtures of experts (ME) model. Our proposed ME model includes a mass point at 0, a mass point at 1, and a mixture of simplex distributions. We implement an EM algorithm in the model fitting. Standard error estimation, model selection and group comparison are also discussed for the ME model. Because of the complexity of the ME model, we also introduce two single model approaches to fit the compliance data. The first approach is the cumulative logit model. The second approach is a quasi-likelihood method. We use data from an asthma medication study to illustrate the proposed model.

In Chapter 4, we extend the methods in Chapter 3 for repeated measures of compliance data. We consider both subject-specific models and the population-averaged models. We introduce a random effects ME model and develop a nonparametric approach in model fitting. A random effects scaled cumulative logit model is discussed in this chapter as well. We also introduce the standard GEE method and the GEE method for repeated ordinal response data. We focus on the study of a marginal ME model, which is called the mixtures of marginal models. We extend the GEE method to the simplex distribution and combine this extension with the ME model to form the mixtures of marginal models. An expectation-solution (ES) algorithm is introduced to fit the mixtures of marginal models. We illustrate and examine the proposed methods in two asthma medication studies.

We summarize this dissertation in the last chapter (Chapter 5) and suggest possible areas for future research.

CHAPTER 2
RANDOM EFFECTS MODELS FOR REPEATED MEASURES OF ZERO-INFLATED COUNT DATA

As we reviewed in the previous chapter, there is considerable literature on modeling cross-sectional zero-inflated count data, but few papers have discussed the modeling of clustered zero-inflated count data. In clustered zero-inflated count data study, besides Hall's (2000) paper and the Yau and Lee (2001) paper, recently Wang, Yau and Lee (2002) used two separate random effects in each part of a ZIP mixed model to analyze clustered zero-inflated count data. The model fitting method in their paper was the PQL method. The above three papers used subject-specific models. For population-averaged studies, Dobbie and Welsh (2001a) applied marginal models using the generalized estimating equations (GEE) approach for both parts of a hurdle model.

In this chapter, we present two types of random effects models for repeated measurements on zero-inflated count data. The first model is a correlated random effects hurdle model, which separately handles the zero observations and the positive counts. In maximum likelihood model fitting, we consider both a normal distribution and a nonparametric approach for the random effects. We also discuss a special form of the hurdle model, which can be used to test the existence of zero-inflation. The second model is a cumulative logit model with random effects, which has the simplicity of using a single model to handle the zero-inflation problem. In the previous chapter we reviewed the cross-sectional hurdle model. Section 2.1 provides a more detailed introduction of the hurdle model with its special cases and uses small simulation studies to discuss the advantages of the hurdle model over the zero-inflated count model. Section 2.2 introduces a hurdle model with random effects and discusses model fitting. The random effects cumulative logit model is introduced in Section 2.3. Section 2.4 illustrates the proposed methods with the data in Yau and Lee (2001). The last section (Section 2.5) discusses the identifiability issue for the ZIP model.

2.1 The Hurdle Model

This section first provides some technical details about ML estimation for the second part of the hurdle model. Then it gives some special cases of the hurdle model, which could be used to test the existence of zero-inflation. At the end of this section, we discuss the reason why we prefer the hurdle model to the zero-inflated count model.

2.1.1 Fitting the Hurdle Model

For response variable Y, let y_i denote the observation for subject i, $i = 1, \ldots, n$. Denote $P(Y_i > 0) = p_i$, and $P(Y_i = 0) = 1 - p_i$. The model assumes that $\{Y_i | Y_i > 0\}$ follow a truncated-at-zero count distribution. Let μ_i be the mean and $g(y_i; \mu_i)$ be the density function for the untruncated count distribution. The resulting distribution has

$$
\begin{aligned}
P(Y_i = 0) &= 1 - p_i, & & (2.1)\\
P(Y_i = j) &= p_i \frac{g(j; \mu_i)}{1 - g(0; \mu_i)}, \qquad j = 1, 2, \ldots. & & (2.2)
\end{aligned}
$$

The corresponding mean of the response variable is

$$
\begin{aligned}
\mathrm{E}(Y_i) &= P(Y_i > 0)\mathrm{E}_g[Y_i | Y_i > 0]\\
&= p_i(\mu_i + \theta_i),
\end{aligned}
$$

where $\mathrm{E}_g[Y_i | Y_i > 0] = \mu_i + \theta_i$, and $\theta_i > 0$ depends on the parameters in the count model. The variance of the response variable is

$$
\mathrm{Var}(Y_i) = p_i \mathrm{Var}_g[Y_i | Y_i > 0] + (1 - p_i)\mathrm{E}_g[Y_i | Y_i > 0].
$$

In general the two parts of the hurdle model allow different explanatory variables. We use a link function $\eta_1[p_i(\boldsymbol{\beta}_1)] = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1$ for the binary process and a link function $\eta_2[\mu_i(\boldsymbol{\beta}_2)] = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$ for the mean of the untruncated count distribution. For instance, the link function for η_1 might be the logistic function, the probit function or the complementary log-log function. In Chapter 1, we used a logistic model for the binary process for illustration. We usually use a log-linear model for the mean of the untruncated count distribution. The likelihood function for the full model is

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n}(1 - p_i(\boldsymbol{\beta}_1))^{I(y_i=0)}\left[p_i(\boldsymbol{\beta}_1)\frac{g\big(y_i; \mu_i(\boldsymbol{\beta}_2)\big)}{1 - g\big(0; \mu_i(\boldsymbol{\beta}_2)\big)}\right]^{1-I(y_i=0)}, \qquad (2.3)$$

where $I()$ is an indicator function. If $1 - p_i > e^{-\mu_i}$ for every i, the model represents zero inflation. If $1 - p_i < e^{-\mu_i}$ for every i, the model represents zero deflation.

As stated in Chapter 1, the log-likelihood factors into two terms,

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \ell_1(\boldsymbol{\beta}_1) + \ell_2(\boldsymbol{\beta}_2).$$

The first $\ell_1(\boldsymbol{\beta}_1)$ can be easily maximized by ML estimation for a binary GLM .

Now we consider ML estimation for ℓ_2 in detail, since it will be used in a later section and it is only discussed in detail for the Poisson case in the current literature. First, suppose that $g(y_i; \mu_i)$ follows a Poisson distribution. Grogger and Carson (1991) discussed this class of models for truncated count data. For it,

$$g(y_i; \mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \dots,$$

and

$$\ell_2(\boldsymbol{\beta}_2) = \sum_{y_i>0}[y_i\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 - e^{\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2} - \log(1 - e^{-\exp(\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2)})] - \sum_{y_i>0}\log(y_i!).$$

We define

$$\ell_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2) = \frac{\partial \ell_2(\boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2} \quad \text{and} \quad \ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2) = \frac{\partial^2 \ell_2(\boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}'_2}.$$

The ML estimator is the solution to

$$\ell_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2) = \sum_{y_i>0}(y_i - \mu_i - \theta_i)\boldsymbol{x}_{2i} = 0,$$

where $\theta_i = \mu_i/(e^{\mu_i} - 1)$. The second derivative is

$$\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2) = -\sum_{y_i>0}(\theta_i + \mu_i)(1 - \theta_i)\boldsymbol{x}_{2i}\boldsymbol{x}'_{2i}.$$

Since the information matrix $I_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2} = -E[\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}] = -\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}$, the Newton-Raphson algorithm and the Fisher scoring algorithm give the same result. Let $\boldsymbol{\beta}_2^{(t)}$ denote the estimate from the t^{th} iteration. We update $\boldsymbol{\beta}_2$ using

$$\boldsymbol{\beta}_2^{(t+1)} = \boldsymbol{\beta}_2^{(t)} - \ell^{-1}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2^{(t)})\ell_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2^{(t)})$$

until the estimate converges.

Next, suppose that $g(y_i;\mu_i)$ is a type II negative binomial distribution (NB2). Then, for dispersion parameter $\alpha > 0$,

$$g(y_i;\mu_i,\alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}.$$

The distribution gives $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i + \alpha\mu_i^2$. As $\alpha \to 0$, the NB2 distribution converges to the Poisson distribution. Since $\log\left(\Gamma(y_i + \alpha^{-1})/\Gamma(\alpha^{-1})\right) = \sum_{l=0}^{y_i-1}\log(l + \alpha^{-1})$, the log-likelihood function is

$$\begin{aligned}
\ell_2(\boldsymbol{\beta}_2, \alpha) &= \sum_{y_i>0}\Bigg[\sum_{l=0}^{y_i-1}\log(l + \alpha^{-1}) + y_i\log(\alpha) + y_i\log(\mu_i) - (y_i + \alpha^{-1})\log(1 + \alpha\mu_i) \\
&\quad - \log(1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}) - \log(y_i!)\Bigg] \\
&= \sum_{y_i>0}\Bigg[\sum_{l=0}^{y_i-1}\log(\alpha l + 1) + y_i\log(\mu_i) - y_i\log(1 + \alpha\mu_i) \\
&\quad - \log((1 + \alpha\mu_i)^{\alpha^{-1}} - 1) - \log(y_i!)\Bigg].
\end{aligned}$$

The ML estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_2$ are the solutions to

$$\ell_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2, \alpha) = \frac{\partial \ell_2(\boldsymbol{\beta}_2, \alpha)}{\partial \boldsymbol{\beta}_2} = \sum_{y_i>0} \left(\frac{y_i - \mu_i - \theta_i^*}{1+\alpha\mu_i} \right) \boldsymbol{x}_{2i} = 0,$$

where

$$\theta_i^* = \mathrm{E}[y_i | y_i > 0] - \mu_i = \frac{\mu_i}{(1+\alpha\mu_i)^{\alpha^{-1}} - 1},$$

and

$$\begin{aligned} \ell_\alpha(\boldsymbol{\beta}_2, \alpha) &= \frac{\partial \ell_2(\boldsymbol{\beta}_2, \alpha)}{\partial \alpha} \\ &= \sum_{y_i>0} \left[\sum_{l=0}^{y_i-1} \frac{l}{\alpha l + 1} - \frac{y_i\mu_i + \alpha^{-1}(\mu_i + \theta_i^*)}{1+\alpha\mu_i} + \frac{(\mu_i + \theta_i^*)\log(1+\alpha\mu_i)}{\alpha^2\mu_i} \right] \\ &= 0. \end{aligned}$$

The second derivatives are

$$\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2, \alpha) = -\sum_{y_i>0} \frac{\alpha y_i \mu_i + \mu_i + \theta_i^* - \theta_i^{*2}(1+\alpha\mu_i)^{\alpha^{-1}}}{(1+\alpha\mu_i)^2} \boldsymbol{x}_{2i}\boldsymbol{x}_{2i}',$$

$$\begin{aligned} \ell_{\alpha\alpha}(\boldsymbol{\beta}_2, \alpha) \;=\; & -\sum_{y_i>0} \left[\sum_{l=0}^{y_i-1} \frac{l^2}{(\alpha l+1)^2} - \frac{y_i\mu_i^2}{(1+\alpha\mu_i)^2} \right. \\ & -\theta_i^*(\mu_i + \theta_i^*) \left(\frac{\alpha^{-1}}{1+\alpha\mu_i} - \frac{\log(1+\alpha\mu_i)}{\alpha^2\mu_i} \right)^2 \\ & \left. -(\mu_i + \theta_i^*) \left(\frac{\alpha^{-1}\mu_i}{(1+\alpha\mu_i)^2} + \frac{2\alpha^{-2}\mu_i}{1+\alpha\mu_i} - \frac{2\log(1+\alpha\mu_i)}{\alpha^3\mu_i} \right) \right], \end{aligned}$$

and

$$\ell_{\boldsymbol{\beta}_2\alpha}(\boldsymbol{\beta}_2, \alpha) = -\sum_{y_i>0} \left[\frac{y_i - \mu_i - \theta_i^*}{(1+\alpha\mu_i)^2}\mu_i - \theta_i^*(\mu_i + \theta_i^*) \left(\frac{\alpha^{-1}}{(1+\alpha\mu_i)^2} - \frac{\log(1+\alpha\mu_i)}{\alpha^2\mu_i(1+\alpha\mu_i)} \right) \right] \boldsymbol{x}_{2i}.$$

Let $(\alpha^{(t)}, \boldsymbol{\beta}_2^{(t)})$ denote the estimates from the t^{th} iteration. The Newton-Raphson algorithm can be applied to obtain the ML estimates by iterating

$$\begin{pmatrix} \boldsymbol{\beta}_2^{(t+1)} \\ \alpha^{(t+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_2^{(t)} \\ \alpha^{(t)} \end{pmatrix} - \begin{pmatrix} \ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) & \ell_{\boldsymbol{\beta}_2\alpha}(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) \\ \ell_{\boldsymbol{\beta}_2\alpha}'(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) & \ell_{\alpha\alpha}(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) \end{pmatrix}^{-1} \begin{pmatrix} \ell_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) \\ \ell_\alpha(\boldsymbol{\beta}_2^{(t)}, \alpha^{(t)}) \end{pmatrix}.$$

To check whether the truncated negative binomial model fits better than the truncated Poisson model, one can test H_0: $\alpha = 0$ against H_1: $\alpha > 0$. Gurmu (1991) proposed a score test for this.

2.1.2 Special Cases

When Mullahy (1986) first proposed the hurdle model for count data, he assumed that the first part of the process is governed by probability mass function g_1, and $\{Y_i | Y_i > 0\}$ is governed by a truncated-at-zero count probability mass function of g_2. The resulting distribution has

$$\begin{aligned} P(Y_i = 0) &= g_1(0; \mu_1), & & (2.4) \\ P(Y_i = j) &= (1 - g_1(0; \mu_1)) \frac{g_2(j; \mu_2)}{1 - g_2(0; \mu_2)}, \quad j = 1, 2, \ldots . & & (2.5) \end{aligned}$$

This model may simplify to a standard generalized linear model when $g_1(.) = g_2(.)$.

Mullahy (1986) considered only the case in which the two parts of the model have the same covariate vectors. In this case, g_1 and g_2 (with means μ_1 and μ_2 respectively) have identical distribution forms, the same overdispersion parameter values, and the same link functions for modeling the means (i.e. $\eta(\mu_1) = \boldsymbol{x}'\boldsymbol{\beta}_1$ and $\eta(\mu_2) = \boldsymbol{x}'\boldsymbol{\beta}_2$). Heilbron (1994) referred to this type of hurdle model as a *compatible two-part model.* When $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, the hurdle model reduces to a standard GLM count model. Otherwise, for covariate x_k, the difference of $\beta_{2k} - \beta_{1k}$ can be used to explain the k^{th} covariate's effect on the zero-inflation. If $\beta_{2k} - \beta_{1k} > 0$, the k^{th} covariate has a positive effect on causing zero-inflation. If $\beta_{2k} - \beta_{1k} < 0$, the k^{th} covariate has a negative effect on causing zero-inflation. By reparameterizing the hurdle model with $\boldsymbol{\beta}_1 = \boldsymbol{\beta}^* + \boldsymbol{\beta}$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}$, Mullahy (1986) proposed a score test H_0: $\boldsymbol{\beta}^* = \boldsymbol{0}$ for testing the standard GLM count model against the hurdle model.

Heilbron (1994) considered some standard distributions for counts, applying a log-linear model for the mean of the distribution, that is: $\log(\mu_{1i}) = \boldsymbol{x}_i'\boldsymbol{\beta}_1$ and

$\log(\mu_{2i}) = \boldsymbol{x}_i'\boldsymbol{\beta}_2$. If g_1 and g_2 are Poisson probability mass functions, it implies that

$$p_i = P(Y_i > 0) = 1 - \exp[-\mu_{1i}(\boldsymbol{\beta}_1)] = 1 - \exp[-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)].$$

Therefore, the first part of the model is equivalent to using a complementary log-log link to model the probability of being positive or not, which is

$$\log\big(-\log(1 - p_i)\big) = \boldsymbol{x}_i'\boldsymbol{\beta}_1. \tag{2.6}$$

If $g_l(\mu_l, \alpha)$ $(l = 1, 2)$ are NB2 distributions as we described in Section 2.1.1, then,

$$p_i = 1 - \big(1 + \alpha\mu_{1i}(\boldsymbol{\beta}_1)\big)^{-1/\alpha}.$$

The first part of the model for the dichotomous variable is

$$\log\{\alpha^{-1}[(1 - p_i)^{-\alpha} - 1]\} = \boldsymbol{x}_i'\boldsymbol{\beta}_1.$$

When $\alpha = 1$, this first part turns to be a logistic model $\text{logit}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_1$.

Next suppose that $g_l(\mu_l, \alpha)$ are type I negative binomial distribution (NB1). Then, for dispersion parameter $\alpha > 0$, the probability mass function is

$$g(y_i; \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1}\mu_i)}{\Gamma(\alpha^{-1}\mu_i)\Gamma(y_i + 1)} \left(\frac{1}{\alpha + 1}\right)^{\alpha^{-1}\mu_i} \left(\frac{\alpha}{\alpha + 1}\right)^{y_i},$$

which has $\text{E}(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \alpha\mu_i$. As $\alpha \to 0$, the NB1 distribution converges to the Poisson distribution. For this distribution, $p_i = 1 - (\alpha + 1)^{-\alpha^{-1}\mu_i}$. The first part of the model for the dichotomous variable is

$$\log\big(-\log(1 - p_i)\big) = \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \log\Big[\frac{1}{\alpha}\log(1 + \alpha)\Big],$$

which is a complementary log-log model with offset if α is known or fixed (Heilbron 1994).

If the two parts of the hurdle model have the same covariates, it is natural to reduce the number of parameters by letting $P(Y_i = 0)$ depend on covariates only through μ_2. Heilbron (1994) proposed a model called the *zero-altered* model, by assuming that $\mu_1 = \gamma_1 \mu_2^{\gamma_2}$ $(\gamma_1 > 0, \gamma_2 \geq 0)$. This implies that $\boldsymbol{\beta}_1 = \log(\gamma_1) + \gamma_2 \boldsymbol{\beta}_2$. Let $\log(\gamma_1) = \gamma_1'$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}$. Through testing if $\gamma_1' = 0$ and $\gamma_2 = 1$, one could test the existence of zero-inflation for the count data. When we set $\gamma_2 = 1$, if $\gamma_1' < 0$, the data are zero-inflated, since it is equivalent to $\mu_1 < \mu_2$. If $\gamma_1' > 0$, the data are zero-deflated. One could use a likelihood-ratio test to conduct these tests. When we set $\gamma_2 = 1$, this model also has the convenient property that $\boldsymbol{\beta}$ can summarize effects overall. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for the general hurdle model one needs to average results from the two components of the model to make an unconditional comparison.

2.1.3 Hurdle Models vs Zero-Inflated Count Models

An alternative approach for modeling zero-inflated count data is the zero-inflated count model proposed by Lambert (1992). This type of model assumes that data are from a mixture of a regular count distribution, such as the Poisson distribution, and a degenerate distribution at zero. As reviewed in Chapter 1, for a zero-inflated Poisson (ZIP) model, it is assumed that

$$Y_i \sim \begin{cases} 0 & \text{with probability } 1 - \pi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } \pi_i, \end{cases}$$

for subject i. The probability distribution has

$$P(Y_i = 0) = 1 - \pi_i + \pi_i e^{-\lambda_i}, \tag{2.7}$$

$$P(Y_i = j) = \pi_i \frac{e^{-\lambda_i}\lambda_i^j}{j!}, \qquad j = 1, 2, \ldots. \tag{2.8}$$

With explanatory variables, the parameters are modeled by

$$\text{logit}(\pi_i) = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2.$$

The EM algorithm or the Newton-Raphson method can be used to obtain the maximum likelihood estimates. Compared with the hurdle model, this model is more complex to fit, as the model components must be fitted simultaneously. The zero-inflated model is only suitable for zero-inflation problems. However, the hurdle model is also suitable for modeling data with fewer zeros than would be expected under standard distribution assumptions. Therefore, when a data set is zero-deflated at some levels of the covariates, the zero-inflated model may fail because that the model is not defined for $\pi_i < 0$. The hurdle model does not have this problem.

To explore whether the zero-inflated count model fails when data are zero-deflated at some levels of the covariates, we conducted two simple simulation studies. We generated data from hurdle Poisson models, in which at some levels of the covariates, $1 - p(\boldsymbol{\beta}_1) < \exp(-\mu(\beta_2))$. Then we used the ZIP model to fit the simulated data sets. The first simulation study used a binary covariate and the second simulation study used a continuous covariate.

2.1.3.1 Simulation study I

The data were simulated form the following hurdle model:

$$\text{logit}(p_i) = 1.5 - 2x_i,$$

$$\log(\mu_i) = 1 - 2x_i.$$

In this study, x_i is a binary variable taking values 0 or 1. When $x_i = 0$, we expect zero-inflation, since $1/(1 + \exp(1.5)) > \exp(-e^1)$. When $x_i = 1$, we expect zero-deflation, since $1/(1 + \exp(-.5)) < \exp(-e^{-1})$. We chose $n = 1000$, where we generated 700 observations from $x_i = 0$ and 300 from $x_i = 1$. Since

$$\frac{700}{1+\exp(1.5)} + \frac{300}{1+\exp(-.5)} > 700\exp(-e^1) + 300\exp(-e^{-1}),$$

the entire data set (ignoring the covariate) tends to be zero-inflated. From the hurdle model the expected number of zeros is

$$\frac{700}{1+\exp(1.5)}+\frac{300}{(1+\exp(-0.5))}=127.7+186.7=314.4.$$

We simulated 1000 data sets from the working model. First we used an ordinary Poisson model $\log(\mu_i)=\beta_0+\beta_1 x_i$ to fit the data. The average estimates are $\hat{\beta}_0=.864$ (S.E. =.029) and $\hat{\beta}_1=-1.665$ (S.E. =.087). The fitted number of zeros using the average Poisson model estimates is

$$700\text{exp}(-e^{0.864})+300\text{exp}(-e^{0.864-1.665})=65.3+191.5=256.8.$$

These data sets tend to contain more zeros than a Poisson distribution can explain. Van den Broek (1995) introduced a score test to test whether the number of zeros is too large for a Poisson distribution. It compares a standard Poisson model with a ZIP model with no covariates in the first part of the model. Jansakul and Hinde (2002) extended this score test to allow the binary part of a ZIP model to have covariates. We used these two tests on our simulated data. From the results of the two tests, all of the 1000 simulated data sets revealed significant evidence of zero inflation.

We then fitted the simulated data sets using the ZIP model and the hurdle model. The results are given in Table 2–1. We show both the mean and the median parameter estimates in the table. When the estimated parameters are highly variable, the medians represent the estimated parameters better than the means. When the variation of the estimated parameters is small, the model generally performs well. As can be seen from the table, the parameters for the hurdle model are well estimated. In the ZIP model fitting, when $x_i=0$, the parameter β_{10} seems to be reasonably estimated. However, when $x_i=1$, the estimated parameter β_{11} has a very large mean (median) standard error, which

Table 2–1: Comparing the estimated parameters of the ZIP model and the hurdle model for the simulated data sets with $n = 1000$ subjects and a binary predictor variable

	β_{10}	β_{11}	β_{20}	β_{21}
Population	1.5	-2	1	-2
ZIP Model				
Mean estimate	1.952	11.070	0.998	-1.789
Median estimate	(1.948)	(12.186)	(0.999)	(-1.787)
S.E. estimate	0.153	3.836	0.027	0.091
Mean S.E.	0.151	512.590	0.027	0.101
Median S.E.	(0.149)	(417.165)	(0.027)	(0.091)
Hurdle Model				
Mean estimate	1.497	-2.000	0.998	-2.028
Median estimate	(1.497)	(-2.002)	(0.999)	(-2.015)
S.E. estimate	0.100	0.153	0.027	0.221
Mean S.E.	0.098	0.154	0.027	0.214
Median S.E.	(0.098)	(0.154)	(0.027)	(0.210)

Note: S.E. estimate is the standard error of the 1000 estimates of parameter. Mean (median) S.E. is the average of (or the median of) the 1000 estimated standard errors. Values in parentheses are the medians.

implies that $\hat{\beta}_{11}$ varies a lot. The ZIP model fails to fit the data sets well. This tells us that even when a test shows significant evidence of zero inflation, the ZIP model may still not be suitable to fit the data.

2.1.3.2 Simulation study II

The second simulation experiment was generated from a hurdle model

$$\text{logit}(p_i) = 3 - 2x_i,$$

$$\log(\mu_i) = 1 - x_i.$$

The covariate x_i takes on 1000 uniformly spaced values between -5 and 5. For x_i in $(-0.27, 1.75)$, $1 - p_i < \exp(-\mu_i)$ and we expect zero-deflation. For x_i in $(-5.00, -0.27)$ and $(1.75, 5.00)$, $1 - p_i > \exp(-\mu_i)$ and we expect zero-inflation. From the hurdle model we proposed, on average, about 35% of the responses are expected to be zero. Using a Poisson model $\log(\mu_i) = \beta_0 - \beta_1 x_i$ to fit these data

Table 2–2: Comparing the estimated parameters of the ZIP model and the hurdle model for the simulated data sets with n = 1000 subjects and a continuous predictor variable.

	β_{10}	β_{11}	β_{20}	β_{21}
Population	3	-2	1	-1
ZIP Model				
Mean estimate	17.688	-5.918	1.010	-0.998
Median estimate	(6.725)	(-2.346)	(1.010)	(-0.998)
S.E. estimate	51.038	16.586	0.022	0.005
Mean S.E.	3.110	1.039	0.021	0.005
Median S.E.	(1.841)	(0.657)	(0.021)	(0.005)
Hurdle Model				
Mean estimate	3.060	-2.036	0.998	-1.000
Median estimate	(3.034)	(-2.019)	(0.999)	(-1.000)
S.E. estimate	0.297	0.169	0.022	0.005
Mean S.E.	0.282	0.162	0.022	0.005
Median S.E.	(0.279)	(0.160)	(0.022)	(0.005)

sets, we got the average estimates $\hat{\beta}_0 = 0.994$ (S.E. =0.021) and $\hat{\beta}_1 = -1.001$ (S.E. =0.005). From this Poisson model, the fitted number of zeros from the average estimated Poisson model is 32.1%. The simulation results are given in Table 2–2. In the first part of the ZIP model, the big differences between the standard errors of the estimated parameters and the mean (median) standard errors of the parameter estimates imply that the parameter estimates for the first part of the ZIP model are not stable. Therefore, in this case, the ZIP model is not trustworthy.

These simulation experiments studied the performance of the ZIP model and the hurdle model when the data sets were simulated from hurdle models. In Section 2.5, we will study the performance of both models when the data sets are generated from the ZIP models.

2.2 Hurdle Models with Random Effects

2.2.1 Model Specification

Now we extend the hurdle model to clustered, correlated counts. Let y_{ij} be observation j ($j = 1, \ldots, t_i$) for subject (or cluster) i ($i = 1, \ldots, n$). Define

$$u_{ij} = \begin{cases} 0, & \text{if } y_{ij} = 0 \\ 1, & \text{if } y_{ij} > 0 \ , \end{cases}$$

and let $p_{ij} = P(y_{ij} > 0)$. Suppose that the positive count response follows a truncated count distribution with probability mass function $g()$ having mean μ_{ij} for the untruncated count distribution. Let $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$ be random effects designed to account for within-subject correlation. Conditional on $\boldsymbol{b}_i$, we assume that

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i}, \tag{2.9}$$

$$\log(\mu_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i}, \tag{2.10}$$

where $\boldsymbol{x}_{kij}$ and $\boldsymbol{z}_{kij}$ ($k = 1, 2$) are covariate vectors pertaining to the fixed effects $\boldsymbol{\beta}_k$ and the random effects $\boldsymbol{b}_{ki}$. In practice, the simple random intercept form of models is often adequate, in which $\boldsymbol{b}_{1i} = b_{1i}$ and $\boldsymbol{b}_{2i} = b_{2i}$ are univariate and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$.

When the response is observed at repeated times, as in longitudinal studies, a high level of a positive response at one time may be positively correlated with a positive outcome at another time. One can tie the two parts of the model together by assuming that the random effects from the two parts are jointly normal and possibly correlated,

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_{i1} \\ \boldsymbol{b}_{i2} \end{pmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix} \right),$$

where $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are unknown positive-definite matrices. Let $\boldsymbol{\psi}$ represent the unknown parameters, $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma})$. The marginal log-likelihood for the hurdle mixed model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\psi}),$$

where

$$\begin{aligned} L_i(\boldsymbol{\psi}) &= \int \left[\prod_{j=1}^{t_i} (1 - p_{ij})^{1-u_{ij}} \left(p_{ij} \frac{g(y_{ij})}{1 - g(0)} \right)^{u_{ij}} \right] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \int [\prod_{j=1}^{t_i} f_1(u_{ij}|\boldsymbol{b}_{1i}) f_2(y_{ij}, u_{ij}|\boldsymbol{b}_{2i})] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned}$$

and $\phi()$ denotes the normal density function for the random effects.

Here we also define a random effects zero-inflated count model. Assume that the response variable Y_{ij} comes from the zero state with probability $1 - \pi_{ij}$, from a count distribution having mean λ_{ij} with probability π_{ij}. Conditional on the random effect $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$, the random effects zero-inflated count model is defined as

$$\text{logit}(\pi_{ij}) = \boldsymbol{x}_{1ij}' \boldsymbol{\beta}_1 + \boldsymbol{z}_{1ij}' \boldsymbol{b}_{1i}, \tag{2.11}$$

$$\log(\lambda_{ij}) = \boldsymbol{x}_{2ij}' \boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}' \boldsymbol{b}_{2i}. \tag{2.12}$$

The marginal log-likelihood for the mixed effects zero-inflated count model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \int \left[\prod_{j=1}^{t_i} (1 - \pi_{ij} + \pi_{ij} e^{-\lambda_{ij}})^{1-u_{ij}} \left(\pi_{ij} g(y_{ij}; \lambda_{ij}) \right)^{u_{ij}} \right] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i.$$

From our previous analysis, in order to use the zero-inflated count model, we have to assume that the data are zero-inflated for every level of the covariates. This is not realistic in practice. Therefore, we are not proposing use in practice of the random effects zero-inflated count model for the repeated measures setting.

If the same covariates affect both the probability of being positive and the level of the count conditional on it being positive, we can extend the zero-altered model to the repeated measures setting. For simplicity, we let $\boldsymbol{b}_i \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ be a subject-specific random effect for both parts of the hurdle model. Conditional on $\boldsymbol{b}_i$, we assume

$$\log\left(1 - \log(1 - p_{ij})\right) = \gamma_1 + \gamma_2(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) + \boldsymbol{b}_i, \qquad (2.13)$$

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{b}_i. \qquad (2.14)$$

We call this the random effects zero-altered model. Through testing whether $\gamma_1 = 0$ with $\gamma_2 = 1$, one can test the existence of zero-inflation for the clustered count data. When we set $\gamma_2 = 1$, if $\gamma_1 < 0$, the data are zero-inflated; If $\gamma_1 > 0$, the data are zero-deflated. As we explained in Section 2.1.2, when we set $\gamma_2 = 1$, we can use $\boldsymbol{\beta}$ to summarize the covariates' effects overall.

2.2.2 ML Model Fitting with Normal Random Effects

To fit the general mixed effects hurdle model, one first obtains the above marginal likelihood by integrating out the random effects. These integrals are analytically intractable, so numerical or stochastic approximation of them is needed. There are many methods to approximate the ML estimate for generalized linear mixed models (McCulloch and Searle 2001, Fahrmeir and Tutz 2001), such as Gauss-Hermite quadrature, the Monte Carlo EM algorithm, Markov chain Monte Carlo (MCMC), penalized quasi likeihood (PQL) and Laplace approximations. The first three methods have the advantage that they converge to the ML estimate as they are applied more finely. Since the simple random intercept form of models is often adequate in practice, we only discuss the case with $\boldsymbol{b}_{1i} = b_{1i}$ and $\boldsymbol{b}_{2i} = b_{2i}$ univariate and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$. With univariate random intercepts, numerical approximation using the Gauss-Hermite quadrature, which approximates the integral by a finite sum, is adequate.

Let

$$f(\boldsymbol{b}_i) = \prod_{j=1}^{t_i} \left[f_1(u_{ij}|\boldsymbol{b}_{1i}) f_2(y_{ij}, u_{ij}|\boldsymbol{b}_{2i})\right].$$

We assume that

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

and let $\boldsymbol{L}$ be the lower triangular Cholesky factor of $\boldsymbol{\Sigma}$. We transform $\boldsymbol{b}_i = \sqrt{2}\boldsymbol{L}\boldsymbol{c}_i$, where $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$. Then, the likelihood function for the i^{th} observation is

$$\begin{aligned} L_i(\boldsymbol{\psi}) &= \int f(\boldsymbol{b}_i)\frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp(-\frac{1}{2}\boldsymbol{b}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_i)d\boldsymbol{b}_i \\ &= \frac{1}{\pi}\int f(\sqrt{2}\boldsymbol{L}\boldsymbol{c}_i)\exp(-\boldsymbol{c}_i^T\boldsymbol{c}_i)d\boldsymbol{c}_i. \end{aligned}$$

The Gauss-Hermite approximation using m quadrature points for each dimension is:

$$L_i^{GH}(\boldsymbol{\psi}) = \sum_{l_1=1}^{m} v_{l_1}^{(1)} \sum_{l_2=1}^{m} v_{l_2}^{(2)} f(c_{l_1}^{(1)}, c_{l_2}^{(2)}),$$

where $v_{l_k}^{(k)} = \pi^{-\frac{1}{2}}w_{l_k}^{(k)}$, and $c_{l_k}^{(k)}$ and $w_{l_k}^{(k)}$ are the node k and weight k $(k = 1, 2)$ of the univariate Gauss-Hermite integration of order m.

To maximize this approximation for the likelihood function, we use an approximate version of the Fisher scoring method (Green 1984, Raudenbush et al. 2000) to obtain $\hat{\boldsymbol{\psi}}$. Let $S(\boldsymbol{\psi})$ denote the score vector, which is approximated as

$$\begin{aligned} S(\boldsymbol{\psi}) &\approx \sum_{i=1}^{n} S_i^{GH}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \frac{\partial \log L_i^{GH}(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} \\ &= \sum_{i=1}^{n} \frac{1}{L_i^{GH}(\boldsymbol{\psi})} \sum_{l_1=1}^{m} v_{l_1}^{(1)} \sum_{l_2=1}^{m} v_{l_2}^{(2)} f(y_i, \boldsymbol{c}^{(l)}; \boldsymbol{\psi}) \frac{\partial \log f(y_i, \boldsymbol{c}^{(l)}; \boldsymbol{\psi})}{\partial\boldsymbol{\psi}}, \end{aligned}$$

where $\boldsymbol{c}^{(l)} = (c_{l_1}^{(1)}, c_{l_2}^{(2)})'$. The Fishing scoring method obtains the ML estimates by iteratively solving the equation $\boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{(t)} + \boldsymbol{I}^{-1}(\boldsymbol{\psi}^{(t)})S(\boldsymbol{\psi}^{(t)})$ until $\boldsymbol{\psi}^{(t)}$ converges, where $\boldsymbol{I} = -E[\sum_{i=1}^{n} \partial^2 \log L_i/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}']$. The second derivatives are usually difficult to calculate. Thus, we used an approximate scoring procedure with $\boldsymbol{I} \approx \sum_{i=1}^{n} S_i(\boldsymbol{\psi})S_i(\boldsymbol{\psi})^T$.

With univariate random effects, one can use the SAS procedure NLMIXED to fit this type of model as well as the random effects zero-inflated count model. SAS NLMIXED uses the adaptive Gauss-Hermite quadrature (Liu and Pierce 1994, Pinheiro and Bates 1995) to approximate the integrals, and the default maximization approach is the quasi-Newton method.

2.2.3 ML Model Fitting with A Nonparametric Approach

The previous section assumed that $\phi(\boldsymbol{b}_i)$ is a bivariate normal probability density function. Since severe misspecification of the random effects distribution could potentially bias parameter estimation, Aitkin (1999) suggested using an unspecified discrete distribution for the random effects. We extend his nonparametric ML (NPML) method in this section for bivariate random effects in the hurdle model.

We assume that $\phi()$ is an unknown discrete distribution with K mass points $\boldsymbol{m} = (\boldsymbol{m}'_1, \ldots, \boldsymbol{m}'_K)'$ and corresponding probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, where $\boldsymbol{m}_k = (m_{1k}, m_{2k})'$, $k = 1, \ldots, K$. The log-likelihood function is

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k [\prod_{j=1}^{t_i} f(y_{ij}; \boldsymbol{\beta}|\boldsymbol{m}_k)],$$

where

$$f(y_{ij}; \boldsymbol{\beta}|\boldsymbol{m}_k) = f_1(u_{ij}; \boldsymbol{\beta}_1|m_{1k}) f_2(y_{ij}, u_{ij}; \boldsymbol{\beta}_2|m_{2k}).$$

This type of finite mixture model can be related to a latent class model (Aitkin and Rubin 1985), which is useful for model fitting. Suppose that d_{ik} is an indicator that represents whether $\boldsymbol{y}_i$ is drawn from the k^{th} latent group, $\sum_k d_{ik} = 1$. Assume that $(\boldsymbol{y}_i, \boldsymbol{d}_i; \boldsymbol{\beta}|\boldsymbol{m}_k)$ are independently distributed with densities

$$\sum_{k=1}^{K} d_{ik} f(\boldsymbol{y}_i; \boldsymbol{\beta}|\boldsymbol{m}_k) = \prod_{k=1}^{K} f(\boldsymbol{y}_i; \boldsymbol{\beta}|\boldsymbol{m}_k)^{d_{ik}},$$

where $f(\boldsymbol{y}_i; \boldsymbol{\beta}|\boldsymbol{m}_k) = \prod_{j=1}^{t_i} f(y_{ij}; \boldsymbol{\beta}|\boldsymbol{m}_k)$. Assume that $(d_{ik}|\boldsymbol{\pi})$ are *i.i.d.* with multinomial distribution $\prod_{k=1}^{K} \pi_k^{d_{ik}}$. Treating $\{d_{ik}\}$ as missing data, the EM algorithm can be used to estimate this finite mixture model. The complete log-likelihood function is

$$\ell_{(c)}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} d_{ik}[\sum_{j=1}^{t_i} \log f(y_{ij};\boldsymbol{\beta}|\boldsymbol{m}_k) + \log \pi_k]. \tag{2.15}$$

In iteration t, the E-step calculates the expectation of the complete log-likelihood; that is,

$$\begin{aligned} \mathrm{E}[\ell_{(c)}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log f(y_{ij};\boldsymbol{\beta}|\boldsymbol{m}_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k \\ &= h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1) + h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2) + h_3(\boldsymbol{\pi}), \end{aligned}$$

where

$$h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log f_1(u_{ij};\boldsymbol{\beta}_1|m_{1k}), \tag{2.16}$$

$$h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log f_2(y_{ij}, u_{ij};\boldsymbol{\beta}_2|m_{2k}), \tag{2.17}$$

$$h_3(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k, \tag{2.18}$$

and

$$w_{ik}^{(t)} = \frac{\pi_k^{(t)}[\prod_{j=1}^{t_i} f(y_{ij};\boldsymbol{\beta}^{(t)}|\boldsymbol{m}_k^{(t)})]}{\sum_{l=1}^{K}\pi_l^{(t)}[\prod_{j=1}^{t_i} f(y_{ij};\boldsymbol{\beta}^{(t)}|\boldsymbol{m}_l^{(t)})]} \tag{2.19}$$

is the posterior mean of d_{ik}. In the M-step, we maximize $\mathrm{E}[\ell_{(c)}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)})]$ with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(t+1)}$. Since $(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$, $(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ and $\boldsymbol{\pi}$ are in three separate terms, we can maximize $h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$, $h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ and $h_3(\boldsymbol{\pi})$ separately. When maximizing with respect to $\boldsymbol{\pi}$, we need to take the constraint $\sum_{k=1}^{K}\pi_k = 1$ into consideration. Solving the equations

$$\frac{\partial}{\partial \pi_k}[h_3(\boldsymbol{\pi}) - \lambda(\sum_{l=1}^{K}\pi_k - 1)] = \frac{\sum_{i=1}^{n} w_{ik}^{(t)}}{\pi_k} - \lambda = 0$$

yields

$$\pi_k^{(t+1)} = \sum_{i=1}^{n} w_{ik}^{(t)}/n. \tag{2.20}$$

The maximization with respect to $(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$ is a weighted version of binomial distribution ML estimators with logit link. Let $\text{logit}(p_{ijk}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + m_{1k}$. We can get $\boldsymbol{\beta}_1^{(t+1)}$ and $\boldsymbol{m}_1^{(t+1)}$ by solving

$$\frac{\partial h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)}{\partial \boldsymbol{\beta}_1} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{k=1}^{K} w_{ik}^{(t)} (u_{ij} - p_{ijk}) \boldsymbol{x}_{1ij} = 0. \tag{2.21}$$

$$\frac{\partial h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)}{\partial m_{1k}} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} w_{ik}^{(t)} (u_{ij} - p_{ijk}) = 0, \qquad k = 1, \ldots, K. \tag{2.22}$$

The maximization with respect to $(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ is a weighted version of truncated Poisson distribution ML estimators or truncated negative binomial distribution ML estimators as discussed in the Section 2.1.1. For example, if the positive counts follow a truncated Poisson distribution, the ML of $(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ can be obtained by solving:

$$\frac{\partial h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2)}{\partial \boldsymbol{\beta}_2} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{k=1}^{K} w_{ik}^{(t)} u_{ij} (y_{ij} - \mu_{ijk} - \theta_{ijk}) \boldsymbol{x}_{1ij} = 0, \tag{2.23}$$

where $\theta_{ijk} = \mu_{ijk}/(\exp(\mu_{ijk}) - 1)$.

$$\frac{\partial h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2)}{\partial m_{2k}} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} w_{ik}^{(t)} u_{ij} (y_{ij} - \mu_{ijk} - \theta_{ijk}) = 0, \qquad k = 1, \ldots, K. \tag{2.24}$$

Convergence of the EM algorithm can be determined by the Euclidean norm of the difference in parameter estimates. In order to avoid a local maximum, trying different starting values is recommended.

2.2.3.1 Standard error estimation

Standard errors of the fixed effects can be obtained by calculating the inverse of the observed information matrix (Louis 1982). Let $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The observed information matrix is

$$F_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}; \boldsymbol{y}) = \mathrm{E}[-\ell_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}; \boldsymbol{y}, \boldsymbol{b}) | \boldsymbol{y}; \hat{\boldsymbol{\psi}}] - \mathrm{Var}[\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi}; \boldsymbol{y}, \boldsymbol{b}) | \boldsymbol{y}; \hat{\boldsymbol{\psi}}], \tag{2.25}$$

where $\ell(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})$ is the complete log-likelihood function.

$$\begin{aligned}\mathrm{Var}[\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}] \approx & \sum_{i=1}^{n}\sum_{k=1}^{K}\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y}_i,\hat{\boldsymbol{m}}_k)l'_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y}_i,\hat{\boldsymbol{m}}_k)\hat{w}_{ik} \\ & +\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{K}\sum_{k'=1}^{K}\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y}_i,\hat{\boldsymbol{m}}_k)\hat{w}_{ik}l'_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y}_{i'},\hat{\boldsymbol{m}}_{k'})\hat{w}_{i'k'},\end{aligned}$$

where $\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y}_i,\hat{\boldsymbol{m}}_k)=\sum_{j=1}^{t_i}\ell_{\boldsymbol{\psi}}(\boldsymbol{\psi};y_{ij},\hat{\boldsymbol{m}}_k)$. In our model,

$$\ell_{\boldsymbol{\beta}_1}(\boldsymbol{\psi};y_{ij},\hat{\boldsymbol{m}}_k)=\sum_{k=1}^{K}(u_{ij}-\hat{p}_{ijk})\boldsymbol{x}_{1ij},$$

and

$$\ell_{\boldsymbol{\beta}_2}(\boldsymbol{\psi};y_{ij},\hat{\boldsymbol{m}}_k)=\sum_{k=1}^{K}u_{ij}(y_{ij}-\hat{\mu}_{ijk}-\hat{\theta}_{ijk})\boldsymbol{x}_{2ij}.$$

The estimated information matrix is

$$\mathrm{E}[-\ell_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}]=\begin{pmatrix}\mathrm{E}[-\ell_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}] & \mathbf{0} \\ \mathbf{0} & \mathrm{E}[-\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}]\end{pmatrix},$$

where

$$\mathrm{E}[-\ell_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}]=\sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K}\hat{w}_{ik}\hat{p}_{ijk}(1-\hat{p}_{ijk})\boldsymbol{x}_{1ij}\boldsymbol{x}'_{1ij},$$

and

$$\mathrm{E}[-\ell_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{b})|\boldsymbol{y};\hat{\boldsymbol{\psi}}]=\sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K}\hat{w}_{ik}u_{ij}(\hat{\theta}_{ijk}+\hat{\mu}_{ijk})(1-\hat{\theta}_{ijk})\boldsymbol{x}_{2ij}\boldsymbol{x}'_{2ij},$$

if we assume that the positive count follows a truncated Poisson distribution. The observed information matrix is usually a by-product of the EM algorithm at convergence.

2.2.3.2 The number of K

For a given choice of the number K of mass points, the estimated maximized log-likelihood is

$$\ell_K(\hat{\boldsymbol{\psi}})=\sum_{i=1}^{n}\log\sum_{k=1}^{K}\hat{\pi}_k[\prod_{j=1}^{t_i}f(y_{ij};\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{m}}_k)],$$

We define the deviance difference, comparing this model to the simpler non-mixture model, by

$$\text{dev}_K = 2[\ell_K(\hat{\boldsymbol{\psi}}) - \ell_1(\hat{\boldsymbol{\psi}})], \tag{2.26}$$

where $\ell_1(\hat{\boldsymbol{\psi}})$ is the estimated maximum log-likelihood function for $\sum_{i=1}^{n} t_i$ independent responses (i.e. $K = 1$). Although this does not give a formal significance test (since the simpler model is on the boundary of the parameter space), the support size of K can be estimated by starting with $K = 2$ and increasing K until the change in the deviance is small.

2.3 Cumulative Logit Models with Random Effects

Saei, Ward, and McGilchrist (1996) suggested grouping the possible count outcomes into K ordered categories and applying an ordinal response model with random effects. Let $Y_{ij,g}$ be the grouped response variable for observation j on subject i. The threshold model for an ordinal response posits an unobservable variable Z, such that one observes $Y_{ij,g} = k$ (i.e., in category k) if Z is between θ_{k-1} and θ_k. Suppose that Z has a cumulative distribution function $G(z - \eta)$, where η is related to explanatory variables by

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i,$$

for a vector $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ of random effects that account for within-subject correlation. Then,

$$P(Y_{ij,g} \leq k) = P(Z \leq \theta_k) = G(\theta_k - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \boldsymbol{z}'_{ij}\boldsymbol{b}_i).$$

The inverse of the *cdf* of G serves as the link function.

In application with zero-inflated count data, one would take the first category to be the 0 outcomes, and then treat each other outcome as a separate category, or group count values together to form the other $K - 1$ categories. Assuming that G is logistic leads to a logit model for the cumulative probabilities with random effects.

Assuming that G is normal leads to a cumulative probit model with random effects. Saei et al. (1996) proposed a cumulative probit model with random effects and used the PQL approach to estimate the parameters. We propose a cumulative logit model with random effects and use parametric ML for model fitting.

The model has the form

$$\text{logit}[P(Y_{ij,g} \leq k; \boldsymbol{x})] = \eta_{ijk} = \theta_k - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \boldsymbol{z}'_{ij}\boldsymbol{b}_i, \quad k = 1, 2, \ldots, K-1. \qquad (2.27)$$

The probability that $Y_{ij,g}$ takes on the value k is

$$\pi_{ij,k} = P(Y_{ij,g} = k) = \frac{1}{1+\exp(-\eta_{ijk})} - \frac{1}{1+\exp(-\eta_{ij,k-1})} \quad k = 1, 2, \ldots, K,$$

where $\eta_{ij0} = -\infty$. For subject i at occasion j, define $y_{ijk} = 1$ if $Y_{ij,g} = k$ $(k = 1, 2, \ldots, K)$ and $y_{ijk} = 0$ otherwise. Then $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijK})'$ is a K-dimensional vector following a multinomial $\prod_{k=1}^{K} \pi_{ij,k}^{y_{ijk}}$ distribution. Let $f(\boldsymbol{y}_{ij}; \boldsymbol{\beta}|\boldsymbol{b}_i)$ be the multinomial probability mass function and ϕ be the multivariate normal density function with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. The marginal log-likelihood for the cumulative logit mixed model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \int \Big[\prod_{j=1}^{t_i} f(\boldsymbol{y}_{ij}; \boldsymbol{\beta}|\boldsymbol{b}_i)\Big] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i.$$

This is similar to the log-likelihood function in Section 2.2. One can use the SAS procedure NLMIXED to fit this model with ML. Hartzel, Agresti and Caffo (2001) provided a nonparametric approach for the random effects in ML model fitting.

This model has the simplicity of a single model to handle the clump at 0 and the positive outcomes. Elements of $\boldsymbol{\beta}$ summarize effects overall, rather than conditional on the response being positive. This is an important advantage. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for general two-part models one needs to

Table 2–3: Injury frequencies in pre-WRATS and post-WRATS intervention

Y	0	1	2	3	4	8
Pre-WRATS	72	38	17	6	3	1
Post-WRATS	108	24	5	0	0	0

average results from the two components of the model to make an unconditional comparison (e.g., to estimate E(Y) for the groups).

2.4 An Occupational Injury Prevention Program Study

This example is from the paper by Yau and Lee (2001). Their paper evaluated the effectiveness of an occupational injury prevention program used in the cleaning services of the studied Australian hospital. This pilot program used Workplace Risk Assessment Teams (WRATS) intervention to attempt to reduce the expected number of manual handling injuries. The WRATS program utilized a workplace risk identification, assessment and control approach to manual handling injury hazard reduction. The data set comprised injury counts from 137 cleaners who were present in pre-WRATS and post-WRATS intervention. Table 2–3 shows the injury frequencies in pre-WRATS and post-WRATS intervention.

More than 65% of the observations are zero. Yau and Lee conducted tests of overdispersion assuming Poisson distributions (Böhning et al. 1997) separately for pre-WRATS and post-WRATS counts. They found that there is strong evidence of overdispersion for the pre-WRATS period. The explanatory variables include *Time* (0 = pre-WRATS period, 1 = post-WRATS period), *Age*, and *Gender* (0 = female, 1 = male). In order to compare our model fitting with theirs, we incorporate the time of exposure (variable *Exposure*) as an offset for both parts of the model.

2.4.1 Random Effects Hurdle Models

Yau and Lee fitted the two components of their random effects hurdle model separately. This corresponds to fitting our random effects hurdle model with $\rho = 0$. However, they used the PQL approach for model fitting. The PQL estimates are

Table 2–4: Parameter estimation of the final model comparing Yau and Lee (2001) PQL estimates and our ML estimates ($\rho = 0$)

Parameter	PQL		ML	
	Estimate	S.E.	Estimate	S.E
β_{10} (Intercept)	-0.095	0.210	-0.065	0.234
β_{11} (Time)	-1.221	0.288	-1.407	0.342
β_{12} (Gender)	-0.965	0.398	-1.111	0.462
σ_1^2	0.363	0.367	0.837	0.737
β_{20} (Intercept)	-0.463	0.162	-0.636	0.207
β_{21} (Time)	-1.085	0.464	-1.082	0.465
σ_2^2	0.351	0.263	0.379	0.253

known to be biased and inconsistent when the random effects have large variance and the binomial denominator is small (Breslow and Lin 1995, Jiang 1998). We used ML to fit this model. We used deviance tests for model selection. We selected the same final models as they did, in which

$$\text{logit}(p_{ij}) = \beta_{10} + \beta_{11}\text{Time} + \beta_{12}\text{Gender} + b_{1i} + \log(\text{Exposure}), \tag{2.28}$$

$$\log(\mu_{ij}) = \beta_{20} + \beta_{21}\text{Time} + b_{2i} + \log(\text{Exposure}). \tag{2.29}$$

Table 2–4 compares the PQL estimates and the ML estimates for the final model, assuming the correlation coefficient $\rho = 0$ for the random effects. It shows that the PQL estimates are often quite far from the ML estimates. For instance, the PQL estimate of the variance of the random effect for the logistic part of the model is 0.36, but the ML estimate of it is 0.84. The PQL fixed effects estimates for the logistic part of the model are also quite far from the ML estimates. Overall, the PQL approach is inadequate for approximating ML estimates well for this data set. For the ML fitting, -2(log-likelihood)= $-2\ell(\hat{\psi}) = 477.0$.

It is preferable to fit the two components of the model jointly, allowing correlated random effects. Our ML fitting then yielded an estimated value for ρ that is very close to 1. A simple version of this model sets the correlation equal to 1.0, by letting the random effect for the second part of the model be a constant multiple of the random effect for the first part. That is,

$$\log(\mu_{ij}) = \beta_{20} + \beta_{21}\text{Time} + cb_{1i} + \log(\text{Exposure}),$$

where c is a constant and $b_{1i} \sim N(0, \sigma_1^2)$. The parameter estimates and their standard errors are given in Table 2–5, using ML.

For the GLMM with perfectly correlated normal random effects, the final model has $-2\ell(\hat{\boldsymbol{\psi}}) = 470.2$. Compared with the model with uncorrelated random effects, the deviance decreases by 6.8 and the degrees of freedom are the same. This gives us the evidence of efficiency gains of fitting a correlated random effects model over fitting a separated random effects model. We also fit a correlated random effects model with the second component assuming a truncated negative binomial distribution, which gives $-2\ell(\hat{\boldsymbol{\psi}}) = 470.2$. This is almost the same as the one using the truncated Poisson model and the estimated $\hat{\alpha}$ is close to 0. From the likelihood-ratio test, the random effects Poisson hurdle model is adequate for this data set.

In the NPML analysis, with $K = 1$ support point, $-2\ell_1(\hat{\boldsymbol{\psi}}) = 482.4$; with $K = 2$ points, $-2\ell_2(\hat{\boldsymbol{\psi}}) = 469.8$; with $K = 3$ points, $-2\ell_3(\hat{\boldsymbol{\psi}}) = 469.2$. So the deviance difference $\text{dev}_2 = 482.4 - 469.8 = 12.6$. Compared to $\text{dev}_3 = 13.2$, the deviance difference does not change much. We use $K = 2$ in the final model. The estimated mass points are $\hat{\boldsymbol{m}}_1 = (2.06, 1.13)'$ and $\hat{\boldsymbol{m}}_2 = (-0.42, -0.23)'$, with $\hat{\boldsymbol{\pi}} = (0.17, 0.83)'$. We estimated the variance for the estimated random effect distribution by

$$\widehat{\text{Var}}(\boldsymbol{b}) = \sum_{k=1}^{K} \hat{\pi}_k(\hat{\boldsymbol{m}}_k - \hat{\text{E}}(\boldsymbol{b}))(\hat{\boldsymbol{m}}_k - \hat{\text{E}}(\boldsymbol{b}))',$$

where $\hat{\text{E}}(\boldsymbol{b}) = \sum_{k=1}^{K} \hat{\pi}_k \hat{\boldsymbol{m}}_k$. The estimated standard deviation for b_1 is $\hat{\sigma}_1 = .93$, the estimated standard deviation for b_2 is $\hat{\sigma}_2 = 0.51$, and the estimated correlation $\hat{\rho} = 0.9994$, which again suggests that the two random effects are highly correlated.

Table 2–5: ML parameter estimation of the final model, allowing perfectly correlated random effects

Parameter	ML		NPML	
	Estimate	S.E.	Estimate	S.E
β_{10} (Intercept)	-0.050	0.249	0.004	0.245
β_{11}(Time)	-1.479	0.347	-1.428	0.316
β_{12} (Gender)	-1.194	0.482	-1.112	0.418
β_{20} (Intercept)	-0.969	0.275	-0.816	0.214
β_{21}(Time)	-1.187	0.462	-1.239	0.458
c	0.650	0.251	0.548	0.116
σ_1	1.137	0.348	0.934	

Thus, we again used a simpler model with the same random effect for both parts of the model. Table 2–5 shows the estimates.

In this example, the ML fitting and NPML fitting give similar estimated covariate coefficients. In ML fitting of the hurdle model, $\hat{\beta}_{11} = -1.479$ (S.E. = .347) and $\hat{\beta}_{21} = -1.187$ (S.E. = .462). The negative values of the estimated parameters for covariate *Time* tell us that there is significant evidence showing that the WRATS program was effective in reducing both the probability of getting injured and the number of injuries conditional on an injury having happened. In the first part of the hurdle model, the *Gender* effect is also significant. The estimated coefficient is -1.194 with S.E. = 0.275, which implies that the male cleaners had lower probabilities of getting injured than the female cleaners.

We also applied a ZIP mixed model for this data set. However, we found that parameter estimates for the logit component of the ZIP mixed model are not stable. With different starting values, the model gives quite different estimates, but the estimated log-likelihood stays the same. It is questionable whether the data are really zero-inflated and whether the model is identifiable. We fitted a standard Poisson generalized linear mixed model (GLMM) and a standard NB2 GLMM. Both of them have $-2\ell(\hat{\psi})$ of 475.8. The estimated α for the NB2 GLMM

Table 2–6: Parameter estimation for Poisson, NB2 and zero-altered mixed effects models

Parameter	Poisson		NB2		Zero-altered	
	Estimate	S.E.	Estimate	S.E	Estimate	S.E
β_0	-0.717	0.142	-0.717	0.143	-1.01	0.277
β_1 (Time)	-0.973	0.198	-0.973	0.201	-0.916	0.426
β_2 (Gender)	-0.509	0.293	-0.509	0.315	-0.528	0.305
σ	0.650	0.131	0.650	0.137	0.809	.185
α			1.322e-7	0.0003		
γ_1'					0.602	0.813
γ_2					1.205	0.519
$-2\ell(\hat{\psi})$	475.8		475.8		474.0	

approximately equals 0, which means that the data are not overdispersed and the data may be not zero-inflated.

We also fitted a zero-altered mixed model

$$\log\left(1-\log(1-p_{ij})\right) = \gamma_1 + \gamma_2(\beta_0 + \beta_1\text{Time} + \beta_2\text{Gender}) + \log(\text{Exposure}) + b_i,$$

$$\log(\mu_{ij}) = \beta_0 + \beta_1\text{Time} + \beta_2\text{Gender} + \log(\text{Exposure}) + b_i.$$

This model has a $-2\ell(\hat{\psi})$ of 474.0. The results are given in Table 2–6. Compared with the standard Poisson mixed effects model, the difference of the deviances is 1.8 with df=2. Through a likelihood ratio test, this data set is not zero-inflated, which explains why the random effects ZIP model does not perform well.

2.4.2 Random Effects Cumulative Logit Models

This example has only 6 possible outcomes, and only 4 observations had counts of at least 4 (see Table 2–3). Thus, to use the cumulative logit model approach, we grouped the response variable into 5 categories (0, 1, 2, 3, $\geq$ 4). Conditional on a random effect b_i, the cumulative logit model is

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - \boldsymbol{x}_{ij}'\boldsymbol{\beta} - b_i, \quad k = 0, 1, 2, 3,$$

Table 2–7: ML estimation of the cumulative logit mixed model

Parameter	ML estimate	S.E.
θ_0	0.016	0.264
θ_1	1.991	0.343
θ_2	3.579	0.498
θ_3	4.685	0.675
β_1 (Time)	-1.700	0.347
β_2 (Gender)	-1.105	0.498
β_3 (log(Exposure))	1.105	0.298
σ	1.276	0.348

where $b_i \sim N(0, \sigma^2)$ accounts for within-subject correlation. Here we treated the time of exposure as an explanatory variable. The final model is

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - \beta_1\text{Time} - \beta_2\text{Gender} - \beta_3 \log(\text{Exposure}) - b_i, \qquad (2.30)$$

with $-2\ell(\hat{\psi}) = 463.7$. Table 2–7 shows the ML estimates.

The random effects cumulative logit model fitting confirms the conclusions we drew in the Section 2.4.1. The estimated $\hat{\beta}_1 = -1.700$ has a standard error of 0.347. This implies that the effect of WRATS program was significant. The estimated odds that the manual handling injuries fall below any fixed category in pre-WRATS intervention are $\exp(-1.700) = 0.18$ times the estimated odds in post-WRATS intervention. The *Gender* effect $\hat{\beta}_2 = -1.105$ (S.E. $= .498$) is also significant. The estimated odds that the manual handling injuries fall below any fixed category for female cleaners are $\exp(-1.105) = 0.33$ times the estimated odds for male cleaners. These estimates are similar to those from the binary part of the hurdle mixed model (see Table 2–5). This is not surprising since the probability modeled there is the first cumulative probability. The significant *Exposure* effect also tells us that the longer the time of exposure, the higher the possibility of getting injured.

Table 2–8: Summary of $-2\ell(\hat{\psi})$ for ML fitting of various models

Model	$-2\ell(\hat{\psi})$	No. of para.
Hurdle mixed model:		
$\rho = 0$	477.0	7
$\rho = 1$	470.2	7
NPML (K=1)	482.4	7
NPML (K=2)	469.8	10
NPML (K=3)	469.2	12
Cumulative logit mixed model:	463.7	8
Poisson mixed model:	475.8	4
NB2 mixed model:	475.8	5
Zero-altered mixed model:	474.0	6

2.4.3 Comparison of Models

Table 2–8 summarizes the estimated log-likelihood values for various fitted models. In this example, the correlation between the random effects from both parts of the hurdle is very high. Through comparing the estimated log-likelihood values of a perfectly correlated random effects hurdle model and an uncorrelated random effects hurdle model, the efficiency gain of fitting a correlated random effects model is clear.

The unstable ML estimates of the random effects ZIP model suggest that before using a mixed effects ZIP model, we should test for the existence of zero inflation. Fitting a zero-altered mixed model is a good way to test for zero-inflation. In this example, we found that the data are not actually zero-inflated. The usefulness of our hurdle model in this example is that it indicates that gender affected only the probability of getting an injury. When injury happened, gender did not affect the number of injuries a person had. We cannot find this out by fitting a standard Poisson GLMM or a standard NB2 GLMM.

The random effects cumulative logit model has the simplicity of using a single model to summarize effects overall. It is easy to fit and easy to interpret. However, since it only fits the grouped data, when we group several count outcomes together to form one category, we cannot use it in predicting the number of injuries.

Therefore, for repeated measures of count responses with many zero outcomes, we should use a zero-altered mixed model to test for zero-inflation. If zero-inflation exists and one wants to know whether the covariates have different effects on zero outcomes and nonzero outcomes, a random effects hurdle model is a good choice. But if one only cares about the covariate effects on the mean outcomes, a random effects cumulative logit model would be a simple method for solving this problem.

2.5 Identifiability of the ZIP Model

From the simulation studies conducted by Lambert (1992), the estimated parameters for the first part of the ZIP model are biased when the sample size is small. Her simulation results also showed that the confidence intervals for the estimated parameters in the first part of the model are not reliable. From the results of using a random effects ZIP model to analyze the Yau and Lee (2001) data, we found that the parameter estimates for the logit component of the ZIP model were not stable. With different starting values, the model gave quite different estimates, but the estimated log-likelihood stayed the same. Cameron and Trivedi (1998) mentioned that the ZIP model works most satisfactorily if the correlation between the variables entering the two parts of the model is small. Common variables entering both parts make it harder to identify their individual roles. Therefore, in this section we study the identifiability of the ZIP model when the two parts have the same covariates. Identification of a model refers to the situation in which only one set of parameter values uniquely maximizes the likelihood. The model is said to be unidentified when more than one set of values provide a maximum.

Table 2–9: Behavior of ZIP coefficients as estimated from 1000 simulated ZIP trials with x being a binary predictor variable

	β_{10}	β_{11}	β_{20}	β_{21}
Population	1.5	-2	1.5	-2
ZIP Model				
Mean estimate	1.461	-1.248	1.500	-2.114
Median estimate	(1.442)	(-1.865)	(1.504)	(-2.036)
S.E. estimate	0.266	2.811	0.057	0.460
Mean S.E.	0.267	55.899	0.054	0.431
Median S.E.	(0.262)	(0.682)	(0.054)	(0.396)

First we conducted a simulation experiment that is similar to Lambert's simulation study. We generated data from a ZIP model with form

$$\text{logit}(p_i) = \beta_{10} + \beta_{11}x_i,$$

$$\log(\mu_i) = \beta_{20} + \beta_{21}x_i.$$

We let $\boldsymbol{\beta}_1 = (1.5, -2)$ and $\boldsymbol{\beta}_2 = (1.5, -2)$. We generated 1000 data sets with each data set having $n = 200$ observations. The covariate x_i is a binary variable that takes value 0 or 1. Among the 200 observations, 100 of them had $x_i = 0$ and the other 100 observations had $x_i = 1$. With these choices, on average, about 51% of the responses y_i were 0 and 22% of the zeros were from a Poisson distribution. Table 2–9 gives the estimated results. The average estimated $\hat{\beta}_{11}$ was rather far from the population parameter $\beta_{11} = -2$. The average standard error for β_{11} was very large, which implies that the estimation for β_{11} is instable. In this case, the median estimates are more reasonable.

In order to study if the simulated data sets were really zero-inflated or not, we used a zero-altered model to test them. The model is

$$\log[-\log(1-p_i)] = \beta'_{10} + \beta'_{11}x_i,$$

$$\log(\mu_i) = \beta'_{20} + \beta'_{21}x_i.$$

Table 2–10: Behavior of estimated coefficients of the ZIP and the hurdle model from simulated ZIP trials with x being a binary predictor variable

	β_{10}	β_{11}	β_{20}	β_{21}
Population	1.5	-2	1.5	-2
ZIP Model				
Mean estimate	1.458	-1.909	1.500	-2.021
S.E. estimate	0.266	0.755	0.054	0.427
Mean S.E.	0.264	0.618	0.057	0.452
Hurdle Model				
Mean estimate	-0.059	-1.788	1.500	-2.106
S.E. estimate	0.101	0.243	0.057	0.460
Mean S.E.	0.092	0.249	0.054	0.452

As we discussed before, if $\beta'_{10} < \beta'_{20}$ and $\beta'_{10} + \beta'_{11} < \beta'_{20} + \beta'_{21}$, the data are zero-inflated. We used the proposed hurdle model to test for zero-inflation for each simulated trial. Through the tests, we found that some of the trials were not really zero-inflated. These trials caused the bias and instability problem in fitting the ZIP mixed model. Deleting these trials, we obtained the results in Table 2–10.

The ZIP model is a special case of a finite mixture Poisson model. It is a mixture of a degenerate distribution at zero with probability $1 - p$ and a regular Poisson distribution with probability p. Without covariates, the class of finite mixtures of Poisson distributions is identifiable (Teicher 1961, Titterington, Smith and Makov 1985). Wang et al. (1996) extended the proof of the identifiability to a finite mixture Poisson model with covariates only in the Poisson part. The finite mixture model with weights depending on the covariates is a special model family, called the *mixtures of experts* (ME) model. Detailed introdution of the ME model will be given in the next chapter. The ME model originated from the neural network literature (Jacobs, Jordan, Nowlan and Hinton 1991, Jordan and Jacobs 1994) and has been widely used in that area. The ZIP model is a special case of the ME model family. Jiang and Tanner (1999) showed that ME models with Poisson components are identifiable. Therefore, the fixed effects ZIP model is identifiable.

For a random effects ZIP model, one needs to obtain a marginal likelihood by integrating out the random effects to fit the model. Since these integrals are analytically intractable, the marginal likelihood does not have a closed-form expression. It may be difficult to prove the identifiability of the random effects ZIP model. We used two simulation studies to study whether this model has an identifiability problem.

The first simulation study had no covariates. A random intercept was assumed for both parts of the ZIP mixed model.

$$\text{logit}(p_{ij}) = \beta_1 + b_{1i},$$

$$\log(\mu_{ij}) = \beta_2 + b_{2i}.$$

The random effects were taken to be normally distributed and correlated,

$$\boldsymbol{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim N(\mathbf{0},\ \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The data were simulated for $\beta_1 = 1$, $\beta_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 0.8$ and $\rho = 0.9$. We generated 1000 data sets. Each data set contains $n = 100$ subjects, and for each subject, there were $t_i = 10$ repeated measures. The estimated results are given in Table 2–11. As can be seen from the table, both the fixed effect parameters and the parameters in the covariance matrix of the random effect were well estimated.

Table 2–11: Model coefficients as estimated from the simulated data sets without covariates

	β_1	β_2	σ_1	σ_2	ρ
Population	1	2	1	0.8	0.9
Mean estimate	0.990	1.985	0.970	0.780	0.900
S.E. estimate	0.146	0.110	0.161	0.100	0.083
Mean S.E.	0.129	0.082	0.133	0.063	0.062

Table 2–12: Model coefficients as estimated from the simulated data sets with a binary covariate

	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	ρ
Population	1.5	-2	1.5	-2	1	0.8	0.9
Mean estimate	1.463	-2.058	1.447	-2.058	0.945	0.734	0.847
S.E. estimate	0.260	0.364	0.194	0.260	0.252	0.173	0.151
Mean S.E.	0.209	0.359	0.116	0.215	0.194	0.081	0.126

The second simulation study for the random effects ZIP used a binary covariate x_i, which takes value 0 and 1.

$$\text{logit}(p_{ij}) = \beta_{10} + \beta_{11} x_{ij} + b_{1i},$$

$$\log(\mu_{ij}) = \beta_{20} + \beta_{21} x_{ij} + b_{2i}.$$

The data were simulated for $\beta_{10} = 1.5$, $\beta_{11} = -2$, $\beta_{20} = 1.5$, $\beta_{21} = -2$, and the covariance matrix of the random effects was the same as the one in the first simulation study. We generated 1000 data sets. Each data set contains $n = 100$ subjects, half of which have $x = 0$ and the other half have $x = 1$. Each subject has $t_i = 10$ repeated measures. From Table 2–12, we can see that both the fixed effect parameters and the parameters in the covariance matrix of the random effects were well estimated.

In these two simulation studies, the random effects ZIP model seemed identifiable. Therefore, if the data are zero-inflated for each evel of the covariates, the

ZIP model may not have an identification problem. But if at some levels of the covariates the data are not zero-inflated, the ZIP model may fail.

CHAPTER 3
MODELING CROSS-SECTIONAL COMPLIANCE DATA

In recent years, increasing attention has been focused on patient compliance in medical studies. United States health care experts conservatively estimated that half of the 1.8 billion prescribed medicines dispensed yearly are not taken as prescribed (Clepper 1992). The contribution of patient compliance to the success of clinical trials is crucial. Poor compliance can seriously reduce the power of the study and cloud interpretation of the study results, while good compliance makes a well-conducted study. Since patient compliance affects sample sizes, cost, power of the study, and the estimation of treatment effects, studying patient compliance is important to the design, analysis and interpretation of clinical trials.

So far, little attention has been paid to specialized appropriate models for compliance data. Thus, in this chapter we discuss how to use parametric models to fit cross-sectional compliance data. Section 3.1 introduces a two-part model to analyze compliance data. One part applies to the discrete masses at 0 and 1, and one part applies to the continuous proportions. Section 3.2 addresses how to apply the mixtures of experts model in compliance data analysis. It includes model introduction, model fitting and model selection. Section 3.3 introduces two methods that fit compliance data using a single model. The last section uses a real data set to illustrate different methods.

3.1 A Two-Part Model

Similar to using the two-part model to analyze semicontinuous data or using the hurdle model to analyze the zero-inflated count data, we propose a two-part model for compliance data. It permits two clumps with positive probability at the extremes and treats the remaining scale as continuous. We group the data into three ordinal categories. Let Z be the grouped response variable,

$$Z = \begin{cases} 1, & \text{if } Y = 0, \text{ with prob. } \pi_1 \\ 2, & \text{if } 0 < Y < 1, \text{ with prob. } \pi_2 \\ 3, & \text{if } Y = 1, \text{ with prob. } \pi_3. \end{cases}$$

We assume the density function for responses between 0% and 100% is $f(y)$ with mean μ and variance v . The expected value of the response variable is

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}(Y|Z)] \\ &= \pi_1 \times 0 + \pi_2 \mathrm{E}[Y|0 < Y < 1] + \pi_3 \times 1 \\ &= \pi_2 \mathrm{E}[Y|Z = 2] + \pi_3 \\ &= \pi_2 \mu + \pi_3. \end{aligned}$$

Similarly,

$$\mathrm{E}[Y^2] = \pi_2 \mathrm{E}[Y^2|Z = 2] + \pi_3.$$

The variance can be shown to be

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{E}[Y^2] - \mathrm{E}[Y]^2 \\ &= \pi_2 \mathrm{E}[Y^2|Z = 2] + \pi_3 - (\pi_2 \mu + \pi_3)^2 \\ &= \pi_2 (v + \mu^2) + \pi_3 - (\pi_2 \mu + \pi_3)^2. \end{aligned}$$

We propose to use a two-part model to fit the compliance data. The first part is a model for the ordinal response Z, which makes a decision about the category to which the response variable belongs. Conditional on it being in the second category, the second part of the model decides the level of the response variable.

The likelihood function is:

$$\begin{aligned} L &= \prod_{y_i=0} \Pr(Y_i=0) \prod_{0<y_i<1} [\Pr(0<Y_i<1) f(y_i|0<y_i<1)] \prod_{y_i=1} \Pr(Y_i=1) \\ &= \prod_{y_i=0} \Pr(Z_i=1) \prod_{0<y_i<1} \Pr(Z_i=2) f(y_i|0<y_i<1) \prod_{y_i=1} \Pr(Z_i=3). \end{aligned}$$

Therefore, the likelihood function can be separated into two parts. The first part involves the model in the first stage,

$$L_1 = \prod_{y_i=0} \Pr(Z_i=1) \prod_{0<y_i<1} \Pr(Z_i=2) \prod_{y_i=1} \Pr(Z_i=3). \tag{3.1}$$

The second part only involves the model in the second stage,

$$L_2 = \prod_{0<y_i<1} f(y_i|0<y_i<1). \tag{3.2}$$

The two parts are independent. Hence, the maximum likelihood estimation is obtained by separately maximizing L_1 and L_2.

This two-part model is a special case of a mixture model. It is a mixture of a degenerate distribution at zero, a degenerate distribution at one, and a continuous density defined between 0 and 1.

3.1.1 Models for the First Stage

Since we grouped the data into three ordinal categories, it is natural to use a model for ordinal responses to fit Z, such as a cumulative logistic model. Assume for the i^{th} subject ($i = 1, \ldots, n$), $\boldsymbol{x}_{1i}$ is a p-dimensional covariate vector corresponding to the response variable Z_i. The model is

$$\text{logit}[P(Z_i \leq k; \boldsymbol{x})] = \theta_k - \boldsymbol{x}'_{1i}\boldsymbol{\gamma}, \qquad k = 1, 2, \tag{3.3}$$

with $(\boldsymbol{\theta}', \boldsymbol{\gamma}')'$ denoting unknown parameters, where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. So, we have

$$P(Z_i = 1) = \frac{\exp(\theta_1 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})}{1 + \exp(\theta_1 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})},$$

$$\begin{aligned} P(Z_i = 2) &= P(Z_i \leq 2) - P(Z_i \leq 1) \\ &= \frac{\exp(\theta_2 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})}{1 + \exp(\theta_2 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})} - \frac{\exp(\theta_1 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})}{1 + \exp(\theta_1 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})}, \end{aligned}$$

and

$$P(Z_i = 3) = \frac{1}{1 + \exp(\theta_2 - \boldsymbol{x}'_{1i}\boldsymbol{\gamma})}.$$

The cumulative logistic model assumes that covariate effects are the same for each cutpoint. McCullagh (1980) used Fisher scoring algorithms to maximize the likelihood function L_1. This can be done by the SAS procedures PROC LOGISTIC or PROC GENMOD. The same effects $\boldsymbol{\gamma}$ for each logit make the model parsimonious. However, when the data are far from this assumption, the cumulative logit model may not be appropriate. PROC LOGISTIC provides a score test to test whether the effects are the same for each cumulative logit against separate effects.

When the cumulative logit model is not sensible to fit the Z, we should perhaps just use a baseline-category logit model to fit the proportions of different categories. Treating $Z = 1$ as the baseline category, the model is

$$\log \frac{P(Z_i = k+1)}{P(Z_i = 1)} = \boldsymbol{x}'_{1i}\boldsymbol{\gamma}_k, \qquad k = 1, 2.$$

Therefore,

$$P(Z_i = 1) = \frac{1}{1 + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_1) + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_2)},$$

$$P(Z_i = 2) = \frac{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_1)}{1 + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_1) + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_2)},$$

and

$$P(Z_i = 3) = \frac{\exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_2)}{1 + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_1) + \exp(\boldsymbol{x}'_{1i}\boldsymbol{\gamma}_2)}.$$

One can maximize the log-likelihood by using the Newton-Raphson method. The SAS procedure PROC LOGISTIC can be used to obtain the ML estimates. Compared to the cumulative logit model, this multinomial logit model has the disadvantage of requiring another set of parameters to describe effects, because it treats the categorization of Z_i as nominal scale.

3.1.2 Models for Continuous Proportions Between (0,1)

Conditional on the response falling in the second category, the second part of the two-part model is used to fit the continuous proportions between 0 and 1. We introduce several parametric models to handle this type of data. Assume that $\boldsymbol{x}_{2i}$ is the covariate vector corresponding to the second part of the model for subject i.

3.1.2.1 Logistic-normal distribution

Aitchison and Shen (1980) proposed using a logistic-normal distribution to analyze d-component compositional data. If we let $d = 2$, we can use it in the continuous proportions problem. Basically, we use a logistic transformation on the continuous proportion confined between zero and one, and assume the transformed data follow a normal distribution. The model can be written as

$$\text{logit}(y_i|z_i = 2) = \boldsymbol{x}_{2i}'\boldsymbol{\beta} + \epsilon_i, \tag{3.4}$$

where ϵ_i is assumed to be distributed as $N(0, \sigma^2)$. This model is easy to fit. One disadvantage of this model is that the nonlinear transformation makes it difficult to interpret effects in terms of the original data. The assumption of a constant variance may also be problematic in some applications.

3.1.2.2 Beta distribution

An alternative approach assumes that $(Y_i|Z = 2)$ follows a beta distribution with parameters λ_{1i} and λ_{2i}. The density function is

$$f(y_i|z_i = 2) = \frac{\Gamma(\lambda_{1i} + \lambda_{2i})}{\Gamma(\lambda_{1i})\Gamma(\lambda_{1i})} y_i^{\lambda_{1i}-1}(1 - y_i)^{\lambda_{2i}-1}, \tag{3.5}$$

where $0 < y_i < 1$ and $0 < \lambda_{1i}, \lambda_{2i}$. The mean is

$$\text{E}(Y_i|Z_i = 2) = \mu_i = \frac{\lambda_{1i}}{\lambda_{1i} + \lambda_{2i}},$$

and

$$\mathrm{Var}(Y_i|Z_i = 2) = \mu_i(1-\mu_i)\rho_i,$$

where

$$\rho_i = \frac{1}{1+\lambda_{1i}+\lambda_{2i}}.$$

The log-likelihood function is:

$$\begin{aligned} \ell &= \sum_{i=1}^{n} \big[\log\Gamma(\lambda_{1i}+\lambda_{2i}) - \log\Gamma(\lambda_{1i}) - \log\Gamma(\lambda_{2i}) \\ &\quad +\lambda_{1i}\log y_i + \lambda_{2i}\log(1-y_i) - \log\big(y_i(1-y_i)\big)\big]. \end{aligned}$$

For simplicity, we assume that parameter $\rho_i = \rho =$ constant. Let $s = \lambda_{1i} + \lambda_{2i} = 1/\rho - 1 =$ constant. Then we use a logit model to fit the mean:

$$\mathrm{logit}(\mu_i) = \boldsymbol{x}_{2i}'\boldsymbol{\beta}. \tag{3.6}$$

It is easy to show that $\lambda_{1i} = s\mu_i$ and $\lambda_{2i} = s(1-\mu_i)$. The likelihood equations are:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \Big[-\psi(\lambda_{1i}) + \psi(\lambda_{2i}) + \log\Big(\frac{y_i}{1-y_i}\Big)\Big] s\mu_i(1-\mu_i)\boldsymbol{x}_{2i} = 0,$$

$$\frac{\partial \ell}{\partial s} = \sum_{i=1}^{n} \big[\psi(s) - \mu_i\psi(\lambda_{1i}) - (1-\mu_i)\psi(\lambda_{2i}) + \mu_i \log y_i + (1-\mu_i)\log(1-y_i)\big] = 0,$$

where $\psi(t) = \frac{\partial}{\partial t}\log\Gamma(t)$ denotes the digamma function. The second derivatives are

$$\begin{aligned} \frac{\partial^2 \ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} &= \sum_{i=1}^{n} \Big\{\big[-\psi'(\lambda_{1i}) - \psi'(\lambda_{2i})\big] s\mu_i(1-\mu_i) \\ &\quad +\big[-\psi(\lambda_{1i}) + \psi(\lambda_{2i}) + \log\big(\frac{y_i}{1-y_i}\big)\big](1-2\mu_i)\Big\} s\mu_i(1-\mu_i)\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}', \\ \frac{\partial^2 \ell}{\partial s^2} &= \sum_{i=1}^{n} \big[\psi'(s) - \mu_i^2\psi'(\lambda_{1i}) - (1-\mu_i)^2\psi'(\lambda_{2i})\big], \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \ell}{\partial\boldsymbol{\beta}\partial s} &= \sum_{i=1}^{n} \Big\{\big[-\mu_i\psi'(\lambda_{1i}) + (1-\mu_i)\psi'(\lambda_{2i})\big] s \\ &\quad +\big[-\psi(\lambda_{1i}) + \psi(\lambda_{2i}) + \log\big(\frac{y_i}{1-y_i}\big)\big]\Big\}\mu_i(1-\mu_i)\boldsymbol{x}_{2i}, \end{aligned}$$

where $\psi'(t) = \frac{\partial^2}{\partial t^2} \log \Gamma(t)$ denotes the trigamma function. We can use the Newton-Raphson algorithm to get the ML estimates.

3.1.2.3 Simplex distribution

Barndorff-Nielsen and Jørgensen (1991) introduced a class of parametric models on the unit simplex that can be applied in studying compositional data. When compositional data have two components, we have continuous proportions data. The simplex distribution for this type of data is denoted as $S^{-}(\mu, \sigma^2)$. The density function of this simplex distribution is defined as

$$f(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2} d(y; \mu)\}, \tag{3.7}$$

for $0 < y < 1$, where

$$d(y; \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},$$

$0 < \mu < 1$, and $\sigma^2 > 0$. The mean and the variance of Y are

$$\mathrm{E}(Y) = \mu,$$

$$\mathrm{Var}(Y) = \mu(1-\mu) - \frac{1}{\sqrt{2}\sigma} \exp\{\frac{1}{\sigma^2\mu^2(1-\mu)^2}\} \times \Gamma\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\},$$

where $\Gamma(a, b) = \int_b^\infty x^{a-1} e^{-x} dx$ is an incomplete gamma function (Song and Tan 2000). From plots of the simplex distribution (Figure 3–1), we can see that with different parameters μ and σ^2, the distribution can be highly skewed or very flat. The parameters μ and σ^2 can be viewed as position and dispersion parameters (Jørgensen 1997).

In the simplex model, $d(y; \mu)$ satisfies the definition of a *regular unit deviance* (Jørgensen 1997):

$$d(y; y) = 0 \qquad \forall y \in (0, 1),$$

$$d(y; \mu) > 0 \qquad \forall y \neq \mu,$$

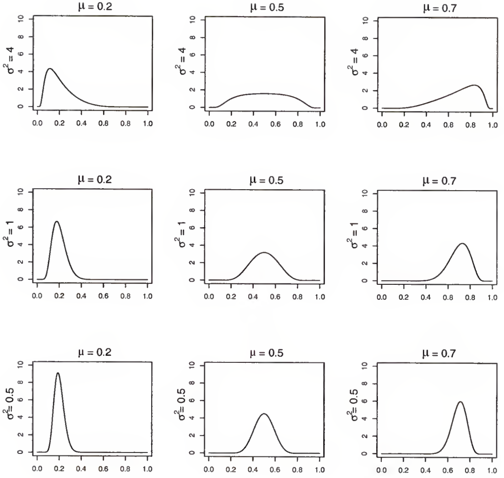


Figure 3–1: Some simplex densities

and

$$\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu) > 0 \qquad \forall \mu \in (0,1).$$

The *unit variance* function V of this regular unit deviance is defined as:

$$V(\mu) = \frac{2}{\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu)} = \mu^3(1-\mu)^3.$$

Therefore, the simplex distribution belongs to *regular proper dispersion models* introduced by Jørgensen (1997), which take the form

$$f(y;\mu,\sigma^2) = a(\sigma^2)V^{-\frac{1}{2}}(y)\mathrm{exp}\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}.$$

where $a()$ is a suitable function, $d(y;u)$ is a regular unit deviance and V is the unit variance function. In the proper dispersion model family, a log-likelihood can be used to construct the unit deviance:

$$d(y;\mu) = c\{l(y;y) - l(y;\mu)\},$$

where c is a constant. The asymptotic variance of Y is $\sigma^2 V(y)$. Exponential dispersion models also belong to the proper distribution model family. Therefore, the simplex distribution shares some common analytic properties with exponential dispersion models.

For the simplex distribution, we use a logit model to fit the mean μ_i,

$$\mathrm{logit}(\mu_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}. \tag{3.8}$$

The log-likelihood function is

$$\ell = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{3}{2}\sum_{i=1}^{n}\log[y_i(1-y_i)] - \frac{1}{2\sigma^2}\sum_{i=1}^{n} d(y_i;\mu_i).$$

The ML estimates for $\boldsymbol{\beta}$ can be obtained by solving the equations

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial d(y_i;\mu_i)}{\partial \boldsymbol{\beta}} = 0.$$

This implies that we can obtain $\hat{\boldsymbol{\beta}}$ by minimizing $d(\boldsymbol{\beta}) = \sum_{i=1}^{n} d(y_i;\mu_i)$. The likelihood equation for estimating σ^2 is

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} d(y_i;\mu_i) = 0.$$

Therefore,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d(y_i;\hat{\mu}_i).$$

We can use the Newton-Raphson algorithm to obtain $\hat{\boldsymbol{\beta}}$. The score function is

$$S(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\frac{\partial d(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)(y_i-2y_i\mu_i+\mu_i^2)}{y_i(1-y_i)\mu_i^2(1-\mu_i)^2}\boldsymbol{x}_{2i},$$

and the second derivative is

$$\ell_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{2y_i^2-(y_i+6y_i^2)\mu_i+(3y_i+6y_i^2)\mu_i^2+(1-6y_i)\mu_i^3+\mu_i^4}{y_i(1-y_i)\mu_i^2(1-\mu_i)^2}\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}'.$$

Let $\boldsymbol{\beta}^{(t)}$ denote the estimate at the t^{th} iteration. We update $\boldsymbol{\beta}$ using

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \ell_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}^{(t)})S(\boldsymbol{\beta}^{(t)})$$

until the estimate converges.

3.2 Mixtures of Experts Models

The two-part model proposed above assumes that the continuous proportions between (0,1) follow a parametric distribution (such as a beta or a simplex distribution). However, in many applications the data do not seem to follow a single distribution, and the response variable is often bimodal or multimodal. For example, by looking at the histograms of our example (Figure 3–2) in Section 3.4, we can see that the response variable obviously does not have a unimodal distribution. The mixtures of experts (ME) model we briefly mentioned in the previous chapter can be used in this situation.

3.2.1 The ME Model

The ME model (Jacobs et al. 1991) originated and has received widespread interest for problems in the area of neural networks. It extends the finite mixture models by using weights depending on covariates. The flexibility feature of modeling has attracted much attention to the study of finite mixtures of distributions. As Mclachlan and Peel (2001) noted, a finite mixture model can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. In the compliance data analysis, we have to deal with the two mass points at 0 and 1, as well as the complex distributional shape between (0, 1). As in the two-part models, the weight for the component to which the response variable belongs also depends on the covariates. The ME models give us more flexibility than the regular finite mixture models.

The ME model is composed of a gating network (the weights) and several expert networks (the components). Let Y_i $(i = 1, \ldots, n)$ denote the response variable for subject i $(i = 1, \ldots, n)$. Let C denote the number of expert networks. For subject i, giving the input (the p-dimensional covariates vector $\boldsymbol{x}_i$), the gating network outputs the weights $(\pi_{ic}(\boldsymbol{x}_i; \boldsymbol{\gamma}_c)$, $c = 1, \ldots, C)$ of the contributions of the expert networks, and the expert networks output the density $f(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}_c)$. The gating network model is modeled by a multinomial logit model.

$$\pi_{ic}(\boldsymbol{x}_i; \boldsymbol{\gamma}_c) = \frac{\exp(\epsilon_{ic})}{\sum_{l=1}^{C} \exp(\epsilon_{il})}, \qquad c = 1, \ldots, C,$$

where $\epsilon_{ic} = \boldsymbol{x}_i' \boldsymbol{\gamma}_c$ with $\boldsymbol{\gamma}_c = (\gamma_{c1}, \ldots, \gamma_{cp}')$.

The parameters in the c^{th} expert network are modeled by

$$h(\theta_c; \boldsymbol{\beta}_c) = \boldsymbol{x}_i' \boldsymbol{\beta}_c,$$

where $h()$ is an appropriate link function and $\boldsymbol{\beta}_c = (\beta_{c1}, \ldots, \beta_{cp})'$.

Set $x_{i1} = 1$ to fit the intercepts in the ME model. Note that the model allows some coefficients to be 0, i.e., $\gamma_{cl} = 0$ and $\beta_{cl'} = 0$ for some l and l', $l, l' = 1, \ldots, p$. Therefore, the probability function of Y_i is:

$$f(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{c=1}^{C} \pi_{ic}(\boldsymbol{x}_i; \boldsymbol{\gamma}_c) f(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}_c).$$

We now extend the ME model to our compliance data study. For subject i, with probability of π_{i0}, the response variable comes from a degenerate distribution of zero; with probability of π_{iC}, the response variable comes from a degenerate distribution of one; with probability of π_{ic} $(c = 1, \ldots, C-1)$ the response variable comes from a density function $f(y_i; \theta_c)$ defined in (0, 1). The density function for the compliance variable is:

$$f(y_i; \boldsymbol{\theta}) = \pi_{i0} I(y_i = 0) + \sum_{c=1}^{C-1} \pi_{ic} f_c(y_i; \theta_c) I(0 < y_i < 1) + \pi_{iC} I(y_i = 1), \qquad (3.9)$$

where $I()$ is the indicator function. For $C = 2$, this is the two-part model we studied in Section 3.1.

For the multinomial logit model we choose $\boldsymbol{\gamma}_0 = 0$ for identifiability. This treats the first category as the baseline category. The model is

$$\log \frac{\pi_{ic}}{\pi_{i0}} = \boldsymbol{x}_i' \boldsymbol{\gamma}_c, \quad c = 1, \ldots, C. \qquad (3.10)$$

For the distribution defined in (0,1), we use the simplex distribution in most of our study. The reasons that we choose the simplex distribution over the other two distributions are as follows: First, the logistic-normal distribution models the transformed responses instead of the original responses, which makes it difficult to interpret. Second, the simplex distribution includes a large class of distributions confined in (0,1). Third, the simplex distribution belongs to the dispersion models (Jørgensen, 1997) and it shares common analytic properties with exponential dispersion models. From Section 3.1.2, we also see that the simplex distribution is easier to fit than the beta distribution. We use a logit model to fit the mean of each simplex distribution. Thus,

$$\text{logit}(\mu_{ic}) = \boldsymbol{x}_i'\boldsymbol{\beta}_c, \quad c = 1, \ldots, C-1, \tag{3.11}$$

where μ_{ic} is the mean of the c^{th} simplex distribution for subject i.

From the model above, the unconditional mean of Y_i is

$$\mathrm{E}(Y_i) = \sum_{c=1}^{C} \pi_{ic}\mu_{ic},$$

where we set $\mu_{iC} = 1$.

3.2.2 Model Fitting

Jordan and Jacobs (1994) introduced the EM algorithm to obtain the ML estimates of the parameters. We implement the EM algorithm to the model we propose. For a fixed number C, let d_{ic} $(c = 0, \ldots, C)$ be an indicator that represents whether Y_i comes from the c^{th} latent group, so $\Pr(d_{ic} = 1) = \pi_{ic}$, and $\sum_{c=0}^{C} d_{ic} = 1$. Thus, $(d_{ic}|\boldsymbol{\pi})$ are *i.i.d.* with multinomial distribution $\prod_{c=0}^{C} \pi_{ic}^{d_{ic}}$. We have

$$d_{i0} = \begin{cases} 1 & \text{if } y_i = 0 \\ 0 & \text{otherwise,} \end{cases}$$

and

$$d_{iC} = \begin{cases} 1 & \text{if } y_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

For $c = 1, \ldots, C-1$, we cannot observe the values of d_{ic}. Therefore, we treat them as missing data. Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$. The log-likelihood for the complete data is

$$\ell_{(c)}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{c=0}^{C} d_{ic}\big[\log f_c(y_i; \boldsymbol{\beta}_c) + \log \pi_{ic}(\boldsymbol{\gamma})\big], \tag{3.12}$$

where for convenience, we assume $f_0(y_i) = 1$; $f_C(y_i) = 1$.

At the $(t+1)^{th}$ E-step, replace the missing data by their expectation conditional on the observed data and the parameter values at the t^{th} step.

$$\begin{aligned}
\mathrm{E}[\ell_{(c)}(\boldsymbol{\psi}^{(t+1)}|\boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^{n}\sum_{c=0}^{C} w_{ic}^{(t)}\left[\log f_c(y_i;\boldsymbol{\beta}_c) + \log \pi_{ic}\right] \\
&= h_1(\boldsymbol{\gamma}) + \sum_{c=1}^{C-1} h_{2c}(\boldsymbol{\beta}_c),
\end{aligned}$$

where

$$h_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{c=0}^{C} w_{ic}^{(t)} \log \pi_{ic}(\boldsymbol{\gamma}), \tag{3.13}$$

$$h_{2c}(\boldsymbol{\beta}_c) = \sum_{i=1}^{n} w_{ic}^{(t)} \log[f_c(y_i;\boldsymbol{\beta}_c)], \quad c = 1,\ldots,C-1, \tag{3.14}$$

$$w_{i0} = d_{i0}, \tag{3.15}$$

$$w_{iC} = d_{iC}, \tag{3.16}$$

and

$$w_{ic}^{(t)} = \frac{\pi_{ic}^{(t)} f_c(y_i;\boldsymbol{\beta}_c^{(t)})}{\sum_{l=1}^{C-1} \pi_{il}^{(t)} f_l(y_i;\boldsymbol{\beta}_l^{(t)})}, \quad c = 1,\ldots,C-1. \tag{3.17}$$

At the $(t+1)^{th}$ M-Step, the expected complete log-likelihood is maximized with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. From the multinomial logit model, we have

$$\pi_{i0} = \frac{1}{1+\sum_{l=1}^{C}\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_l)},$$

and

$$\pi_{ic} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_c)}{1+\sum_{l=1}^{C}\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_l)}, \quad c = 1,\ldots,C.$$

The M-step for $\boldsymbol{\gamma}_c$ $(c = 1,\ldots,C)$ is obtained by solving

$$\frac{\partial h_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_c} = \sum_{i=1}^{n}\sum_{l=0}^{C} w_{il}^{(t)}\left[\frac{1}{\pi_{il}} \times \frac{\partial \pi_{il}}{\partial \epsilon_{ic}} \times \frac{\partial \epsilon_{ic}}{\partial \boldsymbol{\gamma}_c}\right] = 0.$$

Since $\epsilon_{ic} = \boldsymbol{x}_i'\boldsymbol{\gamma}_c$, $\partial\epsilon_{ic}/\partial\boldsymbol{\gamma}_c = \boldsymbol{x}_i$. Chen, Xu and Chi (1999) showed there are three cases of $\partial\pi_{il}/\partial\epsilon_{ic}$:

(a) When $c \neq l$ and $l \neq 0$,

$$\frac{\partial \pi_{il}}{\partial \epsilon_{ic}} = \frac{-\exp(\epsilon_{il})}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} \times \frac{\exp(\epsilon_{ic})}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} = -\pi_{il}\pi_{ic};$$

(b) When $c = l$,

$$\frac{\partial \pi_{il}}{\partial \epsilon_{ic}} = \frac{\exp(\epsilon_{ic})}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} - \frac{\exp(\epsilon_{ic})\exp(\epsilon_{ic})}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} = \pi_{ic}(1 - \pi_{ic});$$

(c) When $l = 0$,

$$\frac{\partial \pi_{il}}{\partial \epsilon_{ic}} = \frac{-1}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} \times \frac{\exp(\epsilon_{ic})}{1 + \sum_{s=1}^{C} \exp(\epsilon_{ic})} = -\pi_{i0}\pi_{ic}.$$

Therefore,

$$\frac{\partial h_1(\gamma)}{\partial \gamma_c} = \sum_{i=1}^{n} \sum_{l=0}^{C} w_{il}^{(t)} (\delta_{cl} - \pi_{ic}) \boldsymbol{x}_i = \sum_{i=1}^{n} (w_{ic}^{(t)} - \pi_{ic}) \boldsymbol{x}_i, \tag{3.18}$$

where δ_{cl} is the Kronecker delta,

$$\delta_{cl} = \begin{cases} 1 & \text{if } c = l \\ 0 & \text{if } c \neq l. \end{cases}$$

Equation (3.18) holds because $\sum_{l=0}^{C} w_{il}^{(t)} = 1$. The second derivatives are

$$\frac{\partial^2 h_1(\gamma)}{\partial \gamma_c \partial \gamma_{c'}'} = -\sum_{i=1}^{n} \pi_{ic}(\delta_{cc'} - \pi_{ic'}) \boldsymbol{x}_i \boldsymbol{x}_i', \quad c, c' = 1, \ldots, C. \tag{3.19}$$

Let

$$J(\gamma) = \frac{\partial h_1(\gamma)}{\partial \gamma} = \left(\frac{\partial h_1(\gamma)}{\partial \gamma_1}, \ldots, \frac{\partial h_1(\gamma)}{\partial \gamma_C} \right)',$$

$$H(\gamma) = \frac{\partial^2 h_1(\gamma)}{\partial \gamma \partial \gamma'} = \begin{bmatrix} H_{11} & H_{12} & \ldots & H_{1C} \\ H_{21} & H_{22} & \ldots & H_{2C} \\ \ldots & \ldots & \ldots & \ldots \\ H_{C1} & H_{C2} & \ldots & H_{CC} \end{bmatrix}.$$

Using the Newton-Raphson method to update the parameter estimate:

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} - H^{-1}(\boldsymbol{\gamma}^{(s)})J(\boldsymbol{\gamma}^{(s)}). \tag{3.20}$$

until convergence occurs.

For $0 < y_i < 1$, calculate the M-step for $h_{2c}(\boldsymbol{\beta}_c)$, $(c = 1, \ldots, C-1)$. The maximization with respect to $\boldsymbol{\beta}_c$ is a weighted version of the ML estimator for the simplex model:

$$J(\boldsymbol{\beta}_c) = \frac{\partial h_{2c}(\boldsymbol{\beta}_c)}{\partial \boldsymbol{\beta}_c} = \sum_{i=1}^{n} w_{ic}^{(t)}\Big(- \frac{1}{2\sigma_c^2}\frac{\partial d(y_i;\mu_{ic})}{\partial \boldsymbol{\beta}_c}\Big), \tag{3.21}$$

and

$$H(\boldsymbol{\beta}_c) = \frac{\partial^2 h_{2c}(\boldsymbol{\beta}_c)}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c'} = \sum_{i=1}^{n} w_{ic}^{(t)}\Big(- \frac{1}{2\sigma_c^2}\frac{\partial^2 d(y_i;\mu_{ic})}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c'}\Big), \tag{3.22}$$

where $\partial d(y_i;\mu_{ic})/\partial \boldsymbol{\beta}_c$ and $\partial^2 d(y_i;\mu_{ic})/\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_c'$ can be found in the previous section. We use the Newton-Raphson method to update the parameter estimates until convergence occurs.

$$\boldsymbol{\beta}_c^{(s+1)} = \boldsymbol{\beta}_c^{(s)} - H^{-1}(\boldsymbol{\beta}_c^{(s)})J(\boldsymbol{\beta}_c^{(s)}). \tag{3.23}$$

Since

$$\frac{\partial h_{2c}(\boldsymbol{\beta}_c)}{\partial \sigma_c^2} = \sum_{i=1}^{n} w_{ic}^{(t)}\big(-\frac{1}{2\sigma_c^2} + \frac{1}{2\sigma_c^4} d(y_i;\mu_{ic})\big),$$

$$\sigma_c^{2(t+1)} = \frac{\sum_{i=1}^{n} w_{ic}^{(t)} d(y_i;\mu_{ic}^{(t+1)})}{\sum_{i=1}^{n} w_{ic}^{(t)}}. \tag{3.24}$$

Therefore, the EM algorithm is summarized as follows:

0. Specify initial values $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ and tolerances values ϵ and ϵ_1.

1. Increase t by 1. (E-Step) For the i^{th} subject, $i = 1, \ldots, n$, compute the weights w_{ic} $(c = 1, \ldots, C-1)$ using (3.17).

2. (M-step) Solve the ML estimators for $\boldsymbol{\gamma}$ in (3.20) using the convergence criterion

$$\frac{||\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^{(t-1)}||}{||\boldsymbol{\gamma}^{(t-1)}||} \le \epsilon_1.$$

Solve the ML estimator for $\boldsymbol{\beta}_c$ ($j = 2, \ldots, C$) in (3.23) using the convergence criterion

$$\frac{||\boldsymbol{\beta}_c^{(t)} - \boldsymbol{\beta}_c^{(t-1)}||}{||\boldsymbol{\beta}_c^{(t-1)}||} \leq \epsilon_1.$$

3. Iterate between 1 and 2 until the overall convergence criteria is satisfied, which is

$$|\ell(\boldsymbol{\psi}^{(t)}) - \ell(\boldsymbol{\psi}^{(t-1)})| \leq \epsilon,$$

where $\ell(\boldsymbol{\psi}^{(t)})$ is the observed log-likelihood function.

3.2.3 Standard error estimation

We use Louis's method (1982) to approximate the observed information matrix. The complete log-likelihood function is

$$\ell_{(c)}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \ell_{i(c)}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{l=0}^{C} d_{il} \ell_{il}(\boldsymbol{\psi}),$$

where

$$\ell_{il}(\boldsymbol{\psi}) = \log f_l(y_i; \boldsymbol{\beta}_l) + \log \pi_{il}(\boldsymbol{\gamma}).$$

The observed variance-covariance matrix is then estimated by

$$\begin{aligned} & \mathrm{E}\left[-\frac{\partial^2 \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}} \right] - \mathrm{Var}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}} \right] & (3.25) \\ = \; & \mathrm{E}\left[-\frac{\partial^2 \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}} \right] - \mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'} \bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}} \right]. & (3.26) \end{aligned}$$

The estimated information matrix is

$$\mathrm{E}\left[-\frac{\partial^2 \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}} \right] = \begin{bmatrix} -H(\hat{\boldsymbol{\gamma}}) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & -\mathrm{H}(\hat{\boldsymbol{\beta}}_1) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & -\mathrm{H}(\hat{\boldsymbol{\beta}}_{\mathrm{C-1}}) \end{bmatrix},$$

where $H(\hat{\gamma})$ and $H(\hat{\boldsymbol{\beta}}_c)$ are the estimated second derivatives using the same equations we use in the M-step.

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{n}\frac{\partial \ell_{i(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\sum_{i=1}^{n}\frac{\partial \ell_{i(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right)\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] \\
&= \sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial \ell_{i(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial \ell_{i(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] \\
&\quad +\sum_{i=1}^{n}\sum_{i\neq i'}\mathrm{E}\left[\frac{\partial \ell_{i(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right]\mathrm{E}\left[\frac{\partial \ell_{i'(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] \\
&= \sum_{i=1}^{n}\mathrm{E}\left[\left(\sum_{l=0}^{C}d_{il}\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\sum_{l'=0}^{C}d_{il'}\frac{\partial \ell_{il'}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right)\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] \\
&\quad +\sum_{i=1}^{n}\sum_{i\neq i'}\left(\sum_{l=0}^{C}\hat{w}_{il}\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\sum_{l'=0}^{C}\hat{w}_{i'l'}\frac{\partial \ell_{i'l'}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right).
\end{aligned}$$

Let us define

$$S_{il}(\boldsymbol{\psi}) = \frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \left(\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_1},\ldots,\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_C},\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_1},\ldots,\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_{C-1}}\right)',$$

where

$$\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_c} = (\delta_{cl}-\pi_{ic})\boldsymbol{x}_i, \quad c=1,\ldots,C$$

and

$$\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_c} = \begin{cases} -\frac{1}{2\sigma_l^2}\frac{\partial d(y_i;\mu_{il})}{\partial \boldsymbol{\beta}_c} & \text{if } l=c \\ 0 & \text{if } l\neq c \end{cases} \qquad c=1,\ldots,C-1.$$

Since $d_{il}d_{il'} = d_{il}$ only when $l=l'$, and $d_{il}d_{il'}=0$ otherwise,

$$\begin{aligned}
&\mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial \ell_{(c)}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right] \\
=&\ \sum_{i=1}^{n}\mathrm{E}\left[\sum_{l=0}^{C}d_{il}\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial \ell_{il}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\middle|\boldsymbol{y},\widehat{\boldsymbol{\psi}}\right]+\sum_{i=1}^{n}\sum_{i\neq i'}\sum_{l=0}^{C}\sum_{l'=0}^{C}\hat{w}_{il}\hat{w}_{i'l'}S_{il}(\widehat{\boldsymbol{\psi}})S'_{i'l'}(\widehat{\boldsymbol{\psi}}) \\
=&\ \sum_{i=1}^{n}\sum_{l=0}^{C}\hat{w}_{il}S_{il}(\widehat{\boldsymbol{\psi}})S'_{il}(\widehat{\boldsymbol{\psi}})+\sum_{i=1}^{n}\sum_{i\neq i'}\sum_{l=0}^{C}\sum_{l'=0}^{C}\hat{w}_{il}\hat{w}_{i'l'}S_{il}(\widehat{\boldsymbol{\psi}})S'_{i'l'}(\widehat{\boldsymbol{\psi}}).
\end{aligned}$$

Then, the standard error estimates can be obtained by inverting the estimated observed information matrix, and taking the square root of diagonal elements.

3.2.4 Model Selection

We discussed above how to fit a ME model when the number of components in the model is known. For inference for the parameters, we can apply the likelihood ratio tests. Typically, we don't know C and we need to infer it from the data. When C is not known, we need to test the null hypothesis $C = c_0$ against the alternative hypothesis $C = c_1$ $(c_1 > c_0)$. Under the null hypothesis, $\pi_{c_0+1}, \pi_{c_0+2}, \ldots, \pi_{c_1}$ are zeros. This implies that $\gamma_{c_0+1}, \gamma_{c_0+2}, \ldots, \gamma_{c_1}$ are all $-\infty$ which lie on the boundary of the parameter space. Therefore, the regularity conditions fail to hold for the likelihood-ratio test statistic to have its usual asymptotic null distribution (Wolfe 1971). We cannot apply the likelihood-ratio test to determine C or conduct inference for the parameters. Wang et al. (1996) proposed model selection criteria Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for a finite mixture Poisson regression model. We use AIC and a small-sample adjustment of AIC called AIC$_c$ in our model selection.

Kullback and Leibler (1951) derived an information measure for measuring the "distance" between two models, which is referred to as the K-L (Kullback-Leibler) distance. Akaike (1973) proposed to use the K-L distance as a fundamental basis for model selection. Let f represent the true model that generated the observed data x. Since f represents the complex reality, one may think that it has an infinite number of parameters. Let $g_i(x|\theta_i)$ represent the i^{th} approximating model for data x given the parameters θ_i. The K-L distance between $f(x)$ and $g_i(x|\theta_i)$ is defined as

$$I(f, g_i) = \int f(x) \log\left(\frac{f(x)}{g_i(x|\theta_i)}\right) dx.$$

This cannot be calculated since we do not have full knowledge of $f(x)$, $g_i(x|\theta_i)$ and the sample space for θ_i. The parameter θ_i is usually unknown and must be estimated from the sample data y. Therefore, the model selection criterion changes from minimizing the K-L distance to minimizing the expected K-L distance

$$\hat{I}(f, g_i) = \int f(x) log\Big(\frac{f(x)}{g_i(x|\hat{\theta}_i(y))}\Big) dx.$$

One would like to choose g_i that minimizes the expected K-L distances $\mathrm{E}_{\hat{\theta}}[\hat{I}(f, g_i)]$. Although this cannot be directly estimated, Akaike (1973) found a relation between the maximized log-likelihood and the relative expected K-L distance. Detailed information about the connection between the K-L distance and AIC criterion is provided by Burnham and Anderson (1998).

Let $\ell(\widehat{\boldsymbol{\psi}})$ be the maximized log-likelihood of the observed data and d be the total number of parameters. AIC (Akaike 1973) is defined as

$$\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\psi}}) + 2d. \tag{3.27}$$

AIC criterion selects the model that has the smallest AIC value. Akaike (1973) showed that for large sample, AIC/2 approximately equals the estimated relative expected K-L distance.

Since the AIC is an asymptotically unbiased estimator of the relative expected K-L distance, when the number of the parameters is relatively large compared to the sample size, AIC may perform poorly. Hurvich and Tsai (1989) derived a bias correction to AIC called AIC_c, which is

$$\mathrm{AIC}_c = -2\ell(\widehat{\boldsymbol{\psi}}) + 2d\Big(\frac{n}{n-d-1}\Big). \tag{3.28}$$

Like the AIC criterion, one chooses the model with the smallest AIC_c value. Burnham and Anderson (1998) suggested to use AIC_c when the ratio n/d is small (say, $n/d < 40$).

As Wang et al. (1996) suggested, we conduct the model selection with two steps. At the first step, we fit the full ME models that contain all possible covariates with different Cs. Then we choose C related to the model with the smallest AIC$_c$ (AIC) value. At the second step, for fixed C, we determine the number of parameters by using likelihood-ratio tests for nested models.

3.2.5 Group Comparison

In compliance studies, we sometimes need to compare the average compliances of several groups. Since the ME model is a finite mixture model, this cannot be done directly. We have to average results from all components and make an unconditional comparison.

Let the covariance matrix for the parameter vector $\boldsymbol{\psi}$ be $\boldsymbol{\Sigma}$. Suppose that we would like to compare two groups – group s with corresponding covariate vector $\boldsymbol{x}_s$ and group t with corresponding covariate vector $\boldsymbol{x}_t$. The average compliances for these two groups are

$$\begin{aligned} \mathrm{E}(Y_k) &= \sum_{l=1}^{C} \pi_{kl}\mu_{kl} \\ &= \sum_{l=1}^{C-1} \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\gamma}_l)}{1+\sum_{c=1}^{C}\exp(\boldsymbol{x}_k'\boldsymbol{\gamma}_c)} \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta}_l)}{1+\exp(\boldsymbol{x}_k'\boldsymbol{\beta}_l)} + \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\gamma}_C)}{1+\sum_{c=1}^{C}\exp(\boldsymbol{x}_k'\boldsymbol{\gamma}_c)}, \end{aligned}$$

where $\mu_{kC} = 1$ and $k = s, t$. The difference of the two groups is $D(\boldsymbol{\psi}) = \mathrm{E}(Y_s) - \mathrm{E}(Y_t)$.

The delta method gives

$$\mathrm{Var}[D(\boldsymbol{\psi})] = \boldsymbol{\phi}'\mathrm{Var}[\boldsymbol{\psi}]\boldsymbol{\phi} = \boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi},$$

where $\boldsymbol{\phi} = (\boldsymbol{\phi}_{\boldsymbol{\gamma}_1}', \ldots, \boldsymbol{\phi}_{\boldsymbol{\gamma}_C}', \boldsymbol{\phi}_{\boldsymbol{\beta}_1}', \ldots, \boldsymbol{\phi}_{\boldsymbol{\beta}_{C-1}}')'$ is the vector of the differentials of $D(\boldsymbol{\psi})$ at $\boldsymbol{\psi}$. The differentials are

$$\boldsymbol{\phi}_{\boldsymbol{\gamma}_c} = \left.\frac{\partial D(\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}_c}\right|_{\boldsymbol{\psi}} = \sum_{l=1}^{C} \pi_{sl}(\delta_{cl} - \pi_{sc})\mu_{sl}\boldsymbol{x}_s - \sum_{l=1}^{C} \pi_{tl}(\delta_{cl} - \pi_{tc})\mu_{tl}\boldsymbol{x}_t,$$

where $c = 1, \ldots, C$.

$$\phi_{\boldsymbol{\beta}_c} = \left.\frac{\partial D(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_c}\right|_{\boldsymbol{\psi}} = \pi_{sc}\mu_{sc}(1-\mu_{sc})\boldsymbol{x}_s - \pi_{tc}\mu_{tc}(1-\mu_{tc})\boldsymbol{x}_t,$$

where $c = 1, \ldots, C-1$.

Therefore, the estimated difference is

$$D(\widehat{\boldsymbol{\psi}}) = \sum_{l=1}^{C} \hat{\pi}_{sl}\hat{\mu}_{sl} - \sum_{l=1}^{C} \hat{\pi}_{tl}\hat{\mu}_{tl}, \tag{3.29}$$

where $\widehat{\boldsymbol{\psi}}$ is the estimated parameter vector. The estimated variance of the $D(\boldsymbol{\psi})$ is

$$\widehat{\mathrm{Var}}[D(\widehat{\boldsymbol{\psi}})] = \widehat{\boldsymbol{\phi}}'\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\phi}}, \tag{3.30}$$

where $\widehat{\boldsymbol{\phi}}$ is the estimated vector of differentials of $D(\boldsymbol{\psi})$ at $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\Sigma}}$ is the inverse of the approximated observed information matrix obtained from Section 3.2. We can calculate the approximate $100(1-\alpha)\%$ confidence interval of the difference by

$$D(\widehat{\boldsymbol{\psi}}) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[D(\widehat{\boldsymbol{\psi}})]}. \tag{3.31}$$

If 0 is included in the confidence interval, then the average compliances of these two groups have no significant difference. If 0 is not included in the confidence interval, then we conclude that the average compliances of these two groups are different, at significance level $1-\alpha$.

3.3 Single-Model Approaches

The inference procedures in the mixture models are not so simple. When comparing different groups that are levels of the explanatory variables, one needs to average results from all components of the model to make an unconditional comparison. If one can use a single model to handle the extreme values (zeros and ones) and the continuous proportional responses together, it makes the inference easier. In this section, we introduce two approaches using a single model to fit the compliance data. One approach is an ordinal threshold model, which was proposed by Saei et al. (1996) in modeling zero-inflated nonnegative data. Another approach is the quasi-likelihood method (Wedderburn 1974).

3.3.1 Ordinal Threshold Models

In applying the ordinal threshold model to compliance data analysis, one needs to group the possible outcomes into K ordered categories. If there are many zeros and ones or it is important to estimate the proportions of 0% compliances and 100% compliances, one can take 0 as the first category and 1 as the K^{th} category, then select $K-3$ cutpoints on the 0 to 1 scale to define the other $K-2$ categories. Let $Y_{i,g}$ be the response variable representing to which category subject i belongs. Let G be a cumulative continuous distribution function. The model assumes that

$$P(Y_{i,g} \leq k) = G(\theta_k - \boldsymbol{x}_i'\boldsymbol{\beta}).$$

In our applications, we mainly assume G is a cumulative standard logistic distribution, so the threshold model is a cumulative logistic model. It has the form

$$\text{logit}[P(Y_{i,g} \leq k)] = \theta_k - \boldsymbol{x}_i'\boldsymbol{\beta}, \quad k = 1, 2, \ldots, K-1. \tag{3.32}$$

It follows that

$$P(Y_{i,g} \leq k) = \frac{\exp(\theta_k - \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\theta_k - \boldsymbol{x}_i'\boldsymbol{\beta})}.$$

Alternatively, we can use a cumulative probit model or a proportional-hazards model to fit the grouped data.

When we apply a cumulative logit model to zero-inflated count data analysis, if there are not many possible outcomes, we can treat each outcome as a separate category. Unlike the count data, there are no obvious cutpoints in grouping compliance data. One needs to group the data on a continuous scale. To explore how to choose the number of groups, we conducted two simple simulation studies.

3.3.1.1 Simulation studies

We wanted to explore how the number of response categories affects the parameter estimators and the powers and the sizes of testing for the effect of the explanatory variables included in the model. We conducted two simple simulation studies. We simulated the data from a logistic distribution with location parameter μ and scale parameter s using the R function **rlogis**. The *cdf* of the distribution is

$$F(y) = \frac{\exp((y-\mu)/s)}{1+\exp((y-\mu)/s)}, \quad -\infty < y < \infty.$$

We set $s = 1$ and let $\mu = \boldsymbol{x}'\boldsymbol{\beta}$.

In the first simulation study, we let $\mu = 0.5x$, where x is a binary variable taking values 0 and 1. We chose the sample size $n = 200$, where we generated 100 observations from $x = 0$ and 100 observations from $x = 1$. We simulated 1000 data sets from the working model. Therefore, the accuracy of the power or the size of the test from this Monte Carlo simulation will be around $\pm\sqrt{p(1-p)/1000}$, where p is the observed power or size. We grouped the simulated data into K ($K = 2, 3, 5, 7, 10, 20$) ordered categories with n/K observations in each category. We used a cumulative logit model to fit the grouped variable Y_g,

$$\text{logit}[P(Y_g \le k)] = \theta_k - \beta x, \quad k = 1, 2, \ldots, K-1.$$

We applied SAS LOGISTIC in model fitting, which gives the estimated parameters and the score test statistic and the likelihood-ratio test statistic for testing $\beta = 0$. Table 3–1 gives the power comparisons of the tests for $\beta = 0$. Table 3–2 gives the mean of the parameter estimates, their standard deviation, and the mean of the standard errors.

These tables suggest that when the grouping size changes from $K = 2$ to $K = 3$, the increase of the powers and the decrease of the standard deviations or the mean standard errors are relatively large. When the grouping size changes from

Table 3–1: Power comparisons of the tests for $\beta = 0$ with different numbers of response categories

No. of	Score test			Likelihood-ratio test		
Response	Nominal α level			Nominal α level		
Categories	0.01	0.05	0.10	0.01	0.05	0.10
2	0.165	0.459	0.577	0.196	0.490	0.608
3	0.242	0.483	0.598	0.242	0.483	0.608
5	0.280	0.507	0.615	0.280	0.507	0.615
7	0.270	0.509	0.620	0.270	0.509	0.620
10	0.276	0.519	0.628	0.286	0.519	0.628
20	0.283	0.519	0.631	0.283	0.519	0.640

Table 3–2: Comparisons of parameter estimates for different numbers of response categories

No. of	Simulation I			simulation II		
Response	$\beta = 0.5$			$\beta = 0$		
Categories	Mean($\hat{\beta}$)	Std. Dev.	Mean(S.E.)	Mean($\hat{\beta}$)	Std. Dev.	Mean(S.E.)
2	0.489	0.284	0.286	0.004	0.295	0.284
3	0.494	0.261	0.262	0.0003	0.264	0.260
5	0.497	0.251	0.253	0.003	0.258	0.251
7	0.499	0.250	0.250	0.001	0.255	0.248
10	0.498	0.248	0.249	0.002	0.253	0.247
20	0.498	0.246	0.248	0.001	0.254	0.246

Note: The accuracy of the estimates for β is around ± 0.018.
The accuracy of the estimated standard errors for β is around ± 0.0002.

$K = 3$ to $K = 5$, the changes of powers and the standard deviations are not as large as the previous ones, but still obvious. Comparing $K = 7, 10, 20$ with $K = 5$, the changes are small.

The second simulation experiment is to study the size of the score test and the likelihood-ratio test for $\beta = 0$, as well as the estimated parameters. In the second simulation study, we let $\mu = 0$. We still used x as a binary covariate taking values 0 and 1. The sample size is $n = 200$ with 100 observations having $x = 0$ and 100 observations having $x = 1$. We simulated 1000 data sets from the working model. We grouped the simulated data the same way as in the previous simulation

Table 3–3: Size comparisons of the tests for $\beta = 0$ with different numbers of response categories

No. of	Score test			Likelihood-ratio test		
Response	Nominal α level			Nominal α level		
Categories	0.01	0.05	0.10	0.01	0.05	0.10
2	0.006	0.082	0.133	0.108	0.184	0.235
3	0.015	0.053	0.105	0.015	0.053	0.110
5	0.013	0.049	0.104	0.013	0.049	0.104
7	0.010	0.054	0.097	0.010	0.054	0.097
10	0.014	0.053	0.099	0.014	0.053	0.099
20	0.015	0.054	0.098	0.020	0.059	0.110

study. Table 3–3 gives the size comparisons of the tests for $\beta = 0$. Table 3–2 gives the mean of the parameter estimates, their standard deviation, and the mean of the standard errors. The decreases in the sizes of the likelihood-ratio tests are relatively large when the grouping size changes from 2 to 3. The sizes of the tests are close for $K = 3, 5, 7, 10, 20$. The standard deviations have relatively large decreases from $K = 2$ to $K = 3$ and from $K = 3$ to $K = 5$.

These two simulation studies suggest that when the number of groups is too small, we will lose some efficiency. We suggest that when grouping a continuous variable, the grouping size should be at least three. Sometimes it is better to have more than five groups. However, it is not necessary to group too many categories, such as $K > 10$. With too many categories, one needs to estimate more parameters and the efficiency gain is small.

3.3.1.2 Score test

When applying the cumulative logit model, one needs to check the assumption that the covariate effects are the same for each cutpoint. Peterson and Harrell (1990) applied the likelihood-ratio test and score test for testing the global proportional odds assumption. Since the likelihood-ratio test requires maximization of likelihood functions under both the null and alternative hypothesis, the score test is preferred. Through a simulation study, Peterson and Harrell (1990) showed that score test gives erroneous or suspicious results when the cross-classification table for the response variable by an explanatory variable contains empty cells at an inner value of Y (i.e. $1 < Y < K$) or when the contingency table is sparse. They also showed that the score test gives suspicious results when the number of observations at one of the levels of the response variable is small relative to the total sample size. Therefore, when we group the data, it is usually desirable to avoid small sample size in individual category of the grouped variable.

Assume that there are p covariates in the model. Let

$$\boldsymbol{\psi} = (\theta_1, \dots, \theta_{K-1}, \beta_{11}, \dots, \beta_{1p}, \dots, \beta_{K-1,1}, \dots, \beta_{K-1,p})',$$

where $(\theta_k, \beta_{k1}, \dots, \beta_{kp})$ are the parameters in the k^{th} binary logit model for $Y_g \leq j$. Let $S(\boldsymbol{\psi})$ be the score vector and $I(\boldsymbol{\psi})$ be the expected information matrix with respect to the parameters $\boldsymbol{\psi}$. The null hypothesis under the global proportional odds assumption is

$$H_0; \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{K-1} = \boldsymbol{\beta}.$$

Under H_0, $\boldsymbol{\psi}_0 = (\theta_1, \dots, \theta_{K-1}, \beta_1, \dots, \beta_p)'$. We can use the efficient score statistic introduced by Rao (1947, 1973), to test the null hypothesis. The score statistic is defined as

$$T = S(\widehat{\boldsymbol{\psi}}_0)' I(\widehat{\boldsymbol{\psi}}_0)^{-1} S(\widehat{\boldsymbol{\psi}}_0). \tag{3.33}$$

It has an asymptotic chi-square distribution with $p(K-2)$ degrees of freedom under the null hypothesis. It is calculated by the SAS procedure PROC LOGISTIC.

McCullagh (1980) noted that for a two-sample problem, the score test for testing the hypothesis $\boldsymbol{\beta} = 0$ is exactly the Wilcoxon sum rank test. For comparing the mean compliances of several treatment groups, if we use the cumulative logit model to fit the ungrouped data, the score test for testing the hypothesis $\boldsymbol{\beta} = 0$ is the nonparametric multiple comparison rank test – the Kruskal Wallis test.

3.3.1.3 Scaled cumulative logit models

In many cases, the global proportional odds assumption does not hold. If this assumption is violated, the model may fit poorly (Agresti 2002). Often lack of fit is due to a dispersion effect. We can apply the scaled cumulative logit model proposed by McCullagh (1980) to model compliance data. This type of model incorporates dispersion effects. For the grouped data,

$$\text{logit}[P(Y_{i,g} \leq k)] = \frac{\theta_k - \boldsymbol{x}'_{1i}\boldsymbol{\beta}}{\exp(\boldsymbol{x}'_{2i}\boldsymbol{\gamma})}, \quad k = 1, 2, \ldots, K-1. \tag{3.34}$$

It follows that

$$P(Y_{i,g} \leq k) = \frac{\exp\left(\frac{\theta_k - \boldsymbol{x}'_{1i}\boldsymbol{\beta}}{\exp(\boldsymbol{x}'_{2i}\boldsymbol{\gamma})}\right)}{1 + \exp\left(\frac{\theta_k - \boldsymbol{x}'_{1i}\boldsymbol{\beta}}{\exp(\boldsymbol{x}'_{2i}\boldsymbol{\gamma})}\right)}.$$

For $\boldsymbol{\gamma} = 0$, this model reduces to the cumulative logit model. In fitting this model, we can apply the Fisher-Scoring method to obtain the ML estimates, as in McCullagh (1980). The SAS procedure PROC NLMIXED can be used to fit this type of model.

3.3.2 Quasi-Likelihood Method

Since no single parametric distribution typically handles the entire compliance data set well, it is hard to use a single model to construct a fully specified likelihood function. The quasi-likelihood method introduced by Wedderburn (1974) allows us to estimate the effects of interest without specifying the underlying distribution of the observations. The example in Wedderburn (1974) used the quasi-likelihood method to fit continuous proportions data that contain 0's as responses. Compliance data generally are the observed proportions that patients' behavior coincides with medical advice. For example, the compliance in the example in Section 3.4 is defined as the number of doses of drug obtained from the pharmacy over the number of doses prescribed. However, the data set contains the proportions but not the number of doses prescribed for each patient. The data are binomial but without information on binomial sample sizes. As in Wedderburn's paper (1974), we can treat the compliance data as pseudo-binomial observations and apply the quasi-likelihood method.

Let y_i denote the response for subject i ($i = 1, \ldots, n$). We assume that $\{Y_i\}$ are independent with mean $E(Y_i) = \mu_i$ and variance $\mathrm{Var}(Y_i) = \phi V(\mu_i)$. The quasi-likelihood function $Q(\mu_i; y_i)$ is defined by

$$\frac{\partial Q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)}. \tag{3.35}$$

It is equivalent to

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt + f(y_i),$$

where $f(y_i)$ is a function of y_i. Wedderburn (1974) showed that the quasi-likelihood function has properties similar to a log likelihood function:

$$\mathrm{E}\left(\frac{\partial Q(\mu_i; y_i)}{\partial \mu_i}\right) = 0$$

$$\mathrm{Var}\left(\frac{\partial Q(\mu_i; y_i)}{\partial \mu_i}\right) = -\mathrm{E}\left(\frac{\partial^2 Q(\mu_i; y_i)}{\partial \mu_i^2}\right) = \frac{1}{\phi V(\mu_i)}.$$

For a one-parameter exponential family, $Q(\mu_i; y_i)$ is the same as the likelihood function.

Assume that the mean μ_i relates to the parameters of interest $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ through $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. The quasi-score function for subject i is defined as

$$\boldsymbol{S}(\boldsymbol{\beta}, y_i) = \frac{\partial Q(\mu_i; y_i)}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{y_i - \mu_i}{\phi V(\mu_i)},$$

above. Therefore, the quasi-score function for complete data $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{y}) &= \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{y_i - \mu_i}{\phi V(\mu_i)} \\ &= \boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})/\phi, \end{aligned}$$

where $\boldsymbol{D} = \partial\boldsymbol{\mu}/\boldsymbol{\beta}$, $\boldsymbol{V} = \text{diag}[V(\mu_1), \ldots, V(\mu_n)]$, and $\boldsymbol{V}^{-1}$ denotes the inverse of the covariance matrix $\boldsymbol{V}$. The maximum quasi-likelihood estimator $\hat{\boldsymbol{\beta}}$ is the solution to $\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{y}) = 0$. Wedderburn (1974) and McCullagh and Nelder (1989) showed that the negative expected value of the second derivative of the quasi-likelihood function is

$$\boldsymbol{I}_{\boldsymbol{\beta}} = -E\left(\frac{\partial^2 Q(\boldsymbol{\mu}; \boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}/\phi.$$

Wedderburn (1974) and McCullagh (1983) proposed using the Fisher-scoring method to obtain the maximum quasi-likelihood estimate. Assume that at the t^{th} iteration, the approximated estimates are $\boldsymbol{\beta}^{(t)}$. Then the $(t+1)^{th}$ approximation is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{D}^{(t)\prime}\boldsymbol{V}^{(t)^{-1}}\boldsymbol{D}^{(t)})^{-1}\boldsymbol{D}^{(t)\prime}\boldsymbol{V}^{(t)^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}).$$

One continues this iteration until convergence occurs.

Under a second-moment assumption, McCullagh (1983) showed that the maximum quasi-likelihood estimate is consistent and asymptotically normal,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow N(\boldsymbol{0}, n\boldsymbol{I}_{\boldsymbol{\beta}}^{-1}) \qquad \text{as} \quad n \rightarrow \infty.$$

He also showed that

$$n^{-\frac{1}{2}}\boldsymbol{S}(\boldsymbol{\beta}; \boldsymbol{y}) \rightarrow N(\boldsymbol{0}, n^{-1}\phi\boldsymbol{I}_{\boldsymbol{\beta}}) \qquad \text{as} \quad n \rightarrow \infty.$$

The estimate of ϕ is obtained by the second moment estimator,

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{X^2}{n-p},$$

where X^2 is the generalized Pearson statistic.

3.3.2.1 Hypothesis tests

Denote the number of parameters by p under H_0 and by q under H_1 where $q < p$. It is also assumed that H_1 is nested in H_0. Let $Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})$ and $Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y})$ denote the maximized quasi-likelihood under H_0 and H_1. McCullagh (1983) showed that under H_0,

$$2(Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})) \to \chi^2_{p-q} \qquad \text{as} \quad n \to \infty.$$

For large samples, this can be used to test a full model against a reduced model.

3.3.2.2 Link functions

For compliance data, link functions for binary responses (for which $0 < \mu < 1$) are natural choices of link functions. Choosing the logistic function $g(\mu) = \log\{\mu/(1-\mu)\}$, or the probit function $g(\mu) = \Phi^{-1}(\mu)$, or the complementary log-log function $g(\mu) = \log\{-\log(1-\mu)\}$ ensures that the predicted means lie between zero and one. McCullagh and Nelder (1989, p. 109) compared these link functions. They found that the logistic and the probit function are almost linearly related over the interval $0.1 \leq \mu \leq 0.9$. It is usually difficult to discriminate between these two functions from the goodness-of-fit point of view. So, here we only discuss using the logistic function and the complementary log-log function.

The logistic link and the complementary log-log link can be considered as coming from the family of link functions given by

$$g^{-1}(\boldsymbol{x}'\boldsymbol{\beta}; \theta) = 1 - [1 + \theta \exp(\boldsymbol{x}'\boldsymbol{\beta})]^{-1/\theta}, \qquad \theta \geq 0.$$

When $\theta = 1$, the model reduces to a logistic model; As $\theta \to 0$, the model converges to a complementary log-log model.

Liang and Zeger (1986) showed that even if the variance function $V()$ is misspecified, the QL estimator $\hat{\boldsymbol{\beta}}$ is still consistent and asymptotically normal as long as the link function and the linear predictor are correct (see also Agresti 2002, Chapter 11). The estimator $\hat{\boldsymbol{\beta}}$ becomes less efficient if the variance function is incorrectly specified. However, if the link function is not correct, the estimator $\hat{\boldsymbol{\beta}}$ may not be consistent. Wu and Lee (2001) proposed a quasi-score test statistic for detecting global lack-of-fit of a link function within a particular family of link functions.

Assume that the variance function is known. Let the maximum quasi likelihood estimator under H_0 be $\hat{\boldsymbol{\beta}}_0$. Let $\hat{\mu}_i(\theta_0) = g^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0; \theta_0)$. The quasi score function for θ under the null hypothesis is defined by

$$\begin{aligned} S_\theta(\theta_0, \hat{\boldsymbol{\beta}}_0) &= \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i(\theta_0)}{\hat{\phi} V[\hat{\mu}_i(\theta_0)]} \right) \left(\frac{\partial \mu_i(\theta_0)}{\partial \theta} \right) && (3.36) \\ &= \boldsymbol{W}'\boldsymbol{Z}. && (3.37) \end{aligned}$$

where $\boldsymbol{W} = (\partial\mu_1(\theta_0)/\partial\theta, \ldots, \partial\mu_n(\theta_0)/\partial\theta)'$, and $\boldsymbol{Z}$ is a vector of $\{\frac{y_i - \hat{\mu}_i(\theta_0)}{\hat{\phi} V[\hat{\mu}_i(\theta_0)]}\}$. Let

$$\hat{I}_{\alpha_2\alpha_1} = \frac{1}{\hat{\phi}} \sum_{i=1}^{n} \frac{1}{V[\hat{\mu}_i(\theta_0)]} \left(\frac{\partial \mu_i(\theta_0)}{\partial \alpha_1} \right) \left(\frac{\partial \mu_i(\theta_0)}{\partial \alpha_2} \right)',$$

where α_1, α_2 can be θ or $\boldsymbol{\beta}$. The quasi-score test statistic is defined by

$$T(\theta) = S_\theta(\theta_0, \hat{\boldsymbol{\beta}}_0)'[\hat{I}_{\theta\theta} - \hat{I}_{\theta\boldsymbol{\beta}} \hat{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \hat{I}_{\boldsymbol{\beta}\theta}]^{-1} S_\theta(\theta_0, \hat{\boldsymbol{\beta}}_0). \qquad (3.38)$$

Under certain conditions, Wu and Lee (2001) showed that $T(\theta)$ converges weakly to a central chi-square variable with degrees-of-freedom $= 1$ under H_0. Given a test size α, one can compute $T(\theta)$ and compare it to $\chi^2_{\alpha,1}$ to detect if the link function is correct or use it to get a confidence interval for plausible values of θ for the link function.

We can simplify the quasi-score test statistic using matrix notations. Let $\boldsymbol{V} = \text{diag}\{1/\hat{\phi} V[\hat{\mu}_i(\theta_0)]\}$ and define $\boldsymbol{D}$ as a $n \times p$ matrix where the cell (i, j) of $\boldsymbol{D}$ is $\partial\mu_i(\theta_0)/\partial\beta_j$. Then, the quasi-score test statistic can be written as

$$T(\theta) = \boldsymbol{Z}'\boldsymbol{W}[(\boldsymbol{V}^{1/2}\boldsymbol{W})'(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{V}^{1/2}\boldsymbol{W})]^{-1}\boldsymbol{W}'\boldsymbol{Z},$$

where $\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{D}[\boldsymbol{D}'\boldsymbol{V}\boldsymbol{D}]^{-1}\boldsymbol{D}'\boldsymbol{V}^{1/2}$.

Table 3–4: Basic information statistics

Medication	Total No.	Mean	Std. Dev.	No. of 0%	No. of 100%
ICS	45	0.600	0.345	4	11
CRO	12	0.506	0.377	1	3
SRT	33	0.698	0.285	0	5

3.4 Application

The example we studied here is to compare the mean compliances of three asthma medications (Sherman et al. 2000). The studied population were children with persistent asthma who were Medicaid recipients. A nurse telephoned identified pharmacies to ask for the patients' refilled histories. The compliance was calculated as the number of doses obtained from the pharmacy over number of doses prescribed. The three medications used to treat the asthma were inhaled corticosteroid (*ICS*), nebulized cromolyn (*CRO*), and slow–release theophylline (*SRT*). In this chapter we studied 90 children who took only one medicine. Some basic statistics for the data are shown in Table 3–4. Figure 3–2 shows the histograms of frequency distribution of the compliances for these three medicines.

3.4.1 Without Other Covariates

First without considering any other covariates, we compared the mean compliances of the three medicines. We treated *SRT* as the baseline drug, variables *ICS* (*ICS* $= 1$ if the patient took ICS, *ICS* $= 0$ otherwise) and *CRO* (*CRO* $= 1$ if the patient took CRO, *CRO* $= 0$ otherwise) as the dummy variables.

3.4.1.1 ME models

We first used a ME model to fit the data. Besides the 0% component and the 100% component, we assumed that there are $C-1$ simplex distribution components in the ME model. For fixed number of components C, the weights of the ME model are modeled by a baseline category model,

$$\log\frac{\pi_{ic}}{\pi_{i0}} = \gamma_{c0} + \gamma_{c1}\text{ICS} + \gamma_{c2}\text{CRO}, \quad c = 1, \ldots, C. \tag{3.39}$$

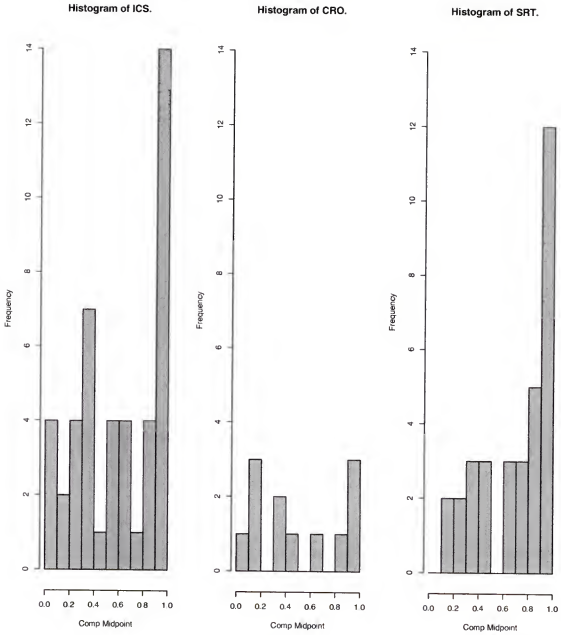


Figure 3–2: Histograms of the frequency distribution of the compliances for the three medicines

Table 3–5: The criterion of selecting C

C	$\ell(\widehat{\boldsymbol{\psi}})$	No. of Para. d	AIC	AIC$_c$
2	-65.616	9	149.232	151.482
3	-45.653	15	121.306	127.306
4	-40.185	21	122.370	135.958

The mean of the c^{th} simplex distribution is modeled by a logit model,

$$\text{logit}(\mu_c) = \beta_{c0} + \beta_{c1}\text{ICS} + \beta_{c2}\text{CRO}, \quad c = 1, \ldots, C-1. \tag{3.40}$$

We chose $C = 2, 3, 4$, and used the model selection criteria AIC and AIC$_c$ to select the number of components. Since there is no count for medicine SRT with 0%, we would have $\hat{\gamma}_{c0} = \infty$ ($c = 1, \ldots, C$). Therefore, we need to add a small amount at this cell to estimate the baseline logit model. We added .1, .01, .001 at this cell. We found that the estimated log-likelihoods changed little, the estimated parameters $\hat{\boldsymbol{\beta}}$ and their standard errors are almost the same, the estimated differeces of $\hat{\gamma}_{c1} - \hat{\gamma}_{c2}$ ($c = 1, 2, 3$) and their standard errors are the same, and the estimated differences of $\hat{\gamma}_{cs} - \hat{\gamma}_{c's}$ ($c, c' = 1, 2, 3$ and $s = 0, 1, 2$) and their standard errors are very close. We Let $\ell(\widehat{\boldsymbol{\psi}})$ be the estimated observed log-likelihood, d be the total number of parameters. Table 3–5 provides the AIC and AIC$_c$ criteria for different Cs when adding .01 to the number of 0% compliance to SRT.

Among the three ME models, both AIC and AIC$_c$ suggest that $C = 3$. This divides the patients into four groups – (a) complete non-compliance group, (b) lower compliance group, (c) high compliance group, and (d) complete compliance group. For $C = 3$, The estimated parameters and their standard errors with adding .01 and .001 to the number of 0% compliance to medicine SRT are shown in Table 3–6. Since the estimated parameters $\hat{\gamma}_{cs}$ ($c = 1, 2, 3$ and $s = 0, 1, 2$) depend on the added small value, we did not report these estimators in the table. We reported

$\gamma_{cs} - \gamma_{3s}$ ($c = 1, 2$ and $s = 0, 1, 2$), which are the estimated parameters of treating the 100% component as the baseline category. The results in the table show that the estimated parameters are robust to the added small amount at the cell with 0 count.

We conducted the Wald tests for γ_{cs}, β_{cs}, $\gamma_{c1} - \gamma_{c2}$, and $\beta_{c1} - \beta_{c2}$ ($c = 1, 2, 3$ and $s = 1, 2$). Only $\hat{\beta}_{21}$ and $\hat{\beta}_{21} - \hat{\beta}_{22}$ are significantly different from zero. The likelihood-ratio statistic for testing the joint effect of the explanatory variables on the weights ($\gamma_{cs} = 0$ for all $c = 1, 2, 3$ and $s = 1, 2$) is 9.16 with df= 6, which has a p-value = 0.164. The likelihood-ratio statistic for testing the joint effect of the explanatory variables on the mean of the first component ($\beta_{11} = \beta_{12} = 0$) is 1.35 with df= 2, which has a p-value = 0.509. The likelihood-ratio statistic for testing the joint effect of the explanatory variables on the mean of the second component ($\beta_{21} = \beta_{22} = 0$) is 8.49 with df= 2, which has a p-value = 0.014. This implies that there is no significant medication effect on the weights of the mixtures of components and the mean of the first simplex component. For patients from the high compliance group, taking ICS ($\hat{\beta}_{21} = 3.701$ with S.E. = 0.467, and $\hat{\beta}_{21} - \hat{\beta}_{22} = 4.503$ with S.E. = 1.004) has a significant effect on increasing the patient's compliance compared to taking SRT and CRO.

Based on the estimated parameters, we calculated the mean compliances using

$$\widehat{\mathrm{E}(Y_i)} = \sum_{c=1}^{C} \hat{\pi}_{ic}\hat{\mu}_{ic}.$$

Table 3–7 gives the estimated means for different components and their estimated weights, as well as the estimated overall mean compliances. The estimated mean compliances for the three medications are close to the observed mean compliances in Table 3–4.

From Table 3–7, we can see that patients who were taking ICS and CRO had larger weights belonging to the non-compliance group and the complete compliance

Table 3–6: Parameter estimation of the ME model (C=3) with adding .01 and .001 to the number of 0% compliance to SRT

Parameter	Adding .01		Adding .001	
	Estimate	S.E.	Estimate	S.E.
$\gamma_{10} - \gamma_{30}$ (Intercept)	-0.132	1.914	-0.132	1.910
$\gamma_{20} - \gamma_{30}$ (Intercept)	1.553	0.595	1.553	0.594
$\gamma_{11} - \gamma_{31}$ (ICS)	1.030	1.947	1.030	1.943
$\gamma_{21} - \gamma_{31}$ (ICS)	-2.852	0.882	-2.852	0.882
$\gamma_{12} - \gamma_{32}$ (CRO)	0.728	2.129	0.728	2.125
$\gamma_{22} - \gamma_{32}$ (CRO)	-1.713	1.635	-1.713	1.635
$\gamma_{11} - \gamma_{12}$ (ICS-CRO)	0.215	1.350	0.215	1.350
$\gamma_{21} - \gamma_{22}$ (ICS-CRO)	-1.226	1.890	-1.226	1.890
$\gamma_{31} - \gamma_{32}$ (ICS-CRO)	-0.087	1.294	-0.087	1.294
β_{10} (Intercept)	-0.413	0.938	-0.413	0.938
β_{11} (ICS-SRT)	0.412	0.950	0.412	0.950
β_{12} (CRO-SRT)	-0.291	1.143	-0.291	1.142
$\beta_{11} - \beta_{12}$ (ICS-CRO)	0.703	0.670	0.703	0.670
β_{20} (Intercept)	0.829	0.401	0.829	0.400
β_{21} (ICS-SRT)	3.701	0.467	3.701	0.466
β_{22} (CRO-SRT)	-0.802	1.054	-0.802	1.054
$\beta_{21} - \beta_{22}$ (ICS-CRO)	4.503	1.004	4.503	1.004
$\ell(\widehat{\psi})$	-45.653		-45.644	

Table 3–7: The estimated mean compliances

Medication	$\hat{\mu}_0$ ($\hat{\pi}_0$)	$\hat{\mu}_1$ ($\hat{\pi}_1$)	$\hat{\mu}_2$ ($\hat{\pi}_2$)	$\hat{\mu}_3$ ($\hat{\pi}_3$)	$\widehat{\mathrm{E}(Y)}$
ICS	0 (0.089)	0.500 (0.600)	0.989 (0.067)	1 (0.244)	0.610
CRO	0 (0.083)	0.331 (0.454)	0.507 (0.213)	1 (0.250)	0.508
SRT	0 (3.03e-4)	0.398 (0.133)	0.696 (0.715)	1 (0.152)	0.702

group than patients who were taking SRT. In group (b) and (c), patients who were taking ICS had much higher mean compliances than patients who were taking CRO and SRT. However, for the ICS patients, the weight belonging to the low compliance group (group (b)) is much larger than the weight belonging to the high compliance group (group (c)) (0.600 vs 0.067), which is the opposite for the SRT patients (0.133 vs 0.715). This explains why the SRT patients had a higher overall mean compliance than that of the ICS patients.

We conducted pairwise comparison for these three medicines using the methods we discussed in Section 3.4. We calculate the approximate 95% Bonferroni confidence intervals for the differences of pairwise comparison (z_α is the $1-\alpha$ quantile of the standard normal distribution):

(1). SRT − ICS:

$$\mathrm{E}(\widehat{Y_{SRT}}) - \mathrm{E}(\widehat{Y_{ICS}}) \pm z_{0.05/6}\mathrm{S.E.} = 0.092 \pm 2.39 \times 0.069 = (-0.073, 0.257).$$

(2). SRT − CRO:

$$\mathrm{E}(\widehat{Y_{SRT}}) - \mathrm{E}(\widehat{Y_{CRO}}) \pm z_{0.05/6}\mathrm{S.E.} = 0.194 \pm 2.39 \times 0.117 = (-0.086, 0.474).$$

(3). ICS − CRO:

$$\mathrm{E}(\widehat{Y_{ICS}}) - \mathrm{E}(\widehat{Y_{CRO}}) \pm z_{0.05/6}\mathrm{S.E.} = 0.102 \pm 2.39 \times 0.117 = (-0.178, 0.382).$$

All the 95% confidence intervals contain 0. Therefore, there is no significant overall mean compliance difference among these three medications, but the confidence intervals are quite wide.

3.4.1.2 Cumulative logit models

Next, we used the cumulative logit model to analyze the data. Since there are no standard rules of how to choose the number of groups and how to choose the cutpoints, we tried several different grouping methods. As we pointed out in

Section 3.3.1, to have a valid score test for the proportional odds assumption, when we group the possible outcomes, we should avoid empty cells at an inner value of Y_g, we should avoid a sparse table, and we should group the data so that the sample size in each category of the grouped variable is sufficiently large.

At this study, a pediatric pulmonologist divided patients into three levels based on his experience: 0% to 50% (1 = very poor), 51% to 84% (2 = less than poor), and 85% to 100% (3 = optimal). Our first grouping method is grouping the data into three categories according to the above criteria. The second grouping method is dividing the data into four ordered categories (0% – 25%, 26% – 50%, 51% – 75%, and 76% – 100%). The third grouping method is dividing the data into five ordered categories (0% – 20%, 21% – 40%, 41% – 60%, 61% – 80% and 81% – 100%). We also tried one method that used 0% and 100% as individual categories, with the rest of the categories as in the second grouping method. Table 3–8 through Table 3–11 give the contingency tables based on these methods.

We fitted the cumulative logit model

$$\text{logit}[P(Y_{i,g} \leq k)] = \theta_k - (\beta_1 \text{ICS} + \beta_2 \text{CRO}), \quad k = 1, 2, \ldots, K-1, \qquad (3.41)$$

using the above four grouping methods. We also fitted the ungrouped data using the cumulative logit model. For the ungrouped data, the score test for testing the hypothesis $\boldsymbol{\beta} = 0$ is the Kruskal Wallis test. Table 3–12 shows the results of different cumulative logit model fittings. In all models based on the four grouping methods, the score tests for the proportional odds assumption are not significant. Therefore, we could use the standard cumulative logit models to fit the grouped data sets. Since the underlying compliance response variable does not follow a logistic distribution, we would not expect that the estimated parameters $\hat{\boldsymbol{\beta}}$ are necessarily close for different grouping methods.

Table 3–8: Cross-classification of medication by categories for grouping method one

Medication	0% – 50%	51% – 84%	84% – 100%
ICS	18	11	16
CRO	7	2	3
SRT	10	10	13

Table 3–9: Cross-classification of medicines by categories for grouping method two

Medication	0% – 25%	26% – 50%	51% – 75%	76% – 100%
ICS	8	10	9	18
CRO	4	3	1	4
SRT	4	6	5	18

Table 3–10: Cross-classification of medicines by categories for grouping method three

Medication	0% – 20%	21% – 40%	41% – 60%	61% – 80%	81% – 100%
ICS	6	11	5	5	18
CRO	4	2	1	1	4
SRT	2	5	3	6	17

Table 3–11: Cross-classification of medicines by categories for grouping method four

Medication	0%	1% – 25%	26% – 50%	51% – 75%	76% – 99%	100%
ICS	4	4	10	9	7	11
CRO	1	3	3	1	1	3
SRT	0	4	6	5	13	5

Table 3–12: Parameter estimation for the cumulative logit models

Parameter	3 categories Estimate (S.E.)	4 categories Estimate (S.E.)	5 categories Estimate (S.E.)
β_1 (ICS-SRT)	-0.286 (0.424)	-0.512 (0.429)	-0.605 (0.428)
β_2 (CRO-SRT)	-0.972 (0.647)	-1.136 (0.621)	-1.242 (0.618)
$\beta_1 - \beta_2$ (ICS-CRO)	0.686 (0.624)	0.624 (0.588)	0.637 (0.583)
T_1 (p-value)	0.728 (0.695) (df = 2)	1.245 (0.871) (df = 4)	3.067 (0.801) (df = 6)
T_2 (p-value)	2.283 (0.319)	3.470 (0.176)	4.319 (0.115)
T_3 (p-value)	2.181 (0.336)	3.289 (0.193)	4.096 (0.129)

Parameter	Ungrouped Estimate (S.E.)	6 categories (Table 3–11) Estimate (S.E.)
β_1 (ICS-SRT)	-0.336 (0.401)	-0.258 (0.406)
β_2 (CRO-SRT)	-0.857 (0.591)	-0.796 (0.600)
$\beta_1 - \beta_2$ (ICS-CRO)	0.521 (0.566)	0.538 (0.576)
T_1 (p-value)	—	12.510 (0.130) (df=8)
T_2 (p-value)	2.069 (0.355)	1.662 (0.436)
T_3 (p-value)	1.939 (0.379)	1.532 (0.465)

Note: T_1 = score test statistic for the proportional odds assumption;
T_2 = likelihood-ratio test statistic for testing $\boldsymbol{\beta} = 0$ (df = 2);
T_3 = score test statistic for testing $\boldsymbol{\beta} = 0$ (df = 2).

Comparing the estimated parameters, $\hat{\beta}_1, \hat{\beta}_2 < 0$ and $\hat{\beta}_1 > \hat{\beta}_2$ for all the models considered. These suggest that the estimated odds that a patient's compliance to medicine SRT falls below any fixed category are lower than the estimated odds of his compliance to ICS and CRO. The estimated odds that a patient's compliance to medicine ICS falls below any fixed category are lower than the estimated odds of his compliance to CRO. The standard errors of the estimated parameters are similar in all the models considered. However, the score tests for testing the null hypothesis $\boldsymbol{\beta} = 0$ are not significant for all the model fittings, so we concluded that the overall compliances for the three asthma medications do not show a significant difference. The purpose of this study is to compare the mean compliances of these three medications. Therefore, it is not necessary to group 0% and 100% as individual categories.

3.4.1.3 Quasi-likelihood method

We next used the quasi-likelihood method to fit this data set. We assume $E(Y_i) = \mu_i$. Since the compliance data can be viewed as pseudo-binomial observations, we first considered the variance function for a binomial distribution $V_1(\mu_i) = \mu_i(1 - \mu_i)$. We also consider variance function $V_2(\mu_i) = [\mu_i(1 - \mu_i)]^2$. This variance function results as an approximation from assuming the response variable follows a logistic-normal distribution. That is, if $\text{logit}(Y) \sim N(\mu, \sigma^2)$, for small σ, $\text{Var}[Y] \approx [\mu(1 - \mu)]^2\sigma^2$.

We used a logit model to model the mean of the response variable,

$$\text{logit}(\mu_i) = \beta_0 + \beta_1\text{ICS} + \beta_2\text{CRO}. \tag{3.42}$$

We let $\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \beta_2)'$ and $\boldsymbol{\beta}_1 = \beta_0$. According to McCullagh (1983), under H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $2(Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y}))$ converges to a χ_2^2 distribution. The fitted values are shown in Table 3–13. The parameter estimators based on these two variance functions are almost the same and the standard errors are close. The

Table 3–13: Parameter estimation for the quasi-likelihood methods

Parameter	$V_1(\mu_i) = \mu_i(1-\mu_i)$ Estimate (S.E.)	$V_2(\mu_i) = [\mu_i(1-\mu_i)]^2$ Estimate (S.E.)
β_0 (Intercept)	0.838 (0.259)	0.838 (0.247)
β_1 (ICS-SRT)	-0.434 (0.332)	-0.434 (0.328)
β_2 (CRO-SRT)	-0.815 (0.417)	-0.815 (0.478)
$\beta_1 - \beta_2$ (ICS-CRO)	0.381 (0.445)	0.381 (0.460)
ϕ	0.682	1.417
$2(Q(\hat{\boldsymbol{\beta}}_1;\boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0;\boldsymbol{y}))$	3.204	3.409
(p-value)	(0.201)	(0.182)

effect estimates $\hat{\beta}_1 = -0.434$ and $\hat{\beta}_2 = -0.815$ suggest that patients taking ICS and CRO had lower mean compliances than patients taking SRT, and $\hat{\beta}_1 > \hat{\beta}_2$ suggests that patients taking CRO had lower mean compliances than patients taking ICS. The test statistics for testing $\beta_1 = \beta_2 = 0$ are not significant. We drew the same conclusion as the ones we had using the other two methods, which is there is no significant overall mean compliance difference among these three medications.

3.4.2 With Other Covariates

In this section, we compared the mean compliances of the three medicines conditional on other covariates. At the beginning of the clinical visit, a pediatric pulmonologist interviewed patients and caretakers, then estimated adherence on a checklist. He assessed the level of the patient's compliance as one of the three levels we defined in the first grouping method for the cumulative logit model. Let the doctor's assessment be the explanatory variable *Assess* (1 = very poor, 2 = less than optimal, and 3 = optimal). The data set also gives each patient's age. There were three patients having missing ages. Two of them were patients who were taking CRO. Both of these patients had a compliance of 100%. Since there were only 12 patients who were taking CRO, if we delete the data with missing covariates, the mean compliance of CRO patients will drop from 0.506 to 0.407. We used Spearman rank correlation statistic (Agresti 2002, p. 90) to examine the correlation between compliance and age. Spearman's correlation statistic is 0.004. Since Spearman's correlation statistic has an asymptotic chi-squared distribution with df=1, it has p-value=.947. This shows that compliance and age do not have a significant correlation. Therefore, we replaced the missing age with the average age for CRO.

3.4.2.1 ME models

First we considered ME models to fit the data. We did not consider the interaction terms in the full ME model because of two reasons. The first reason is that there are only 90 observations, if we add the interaction terms for $C = 4$, the total number of parameter will be 70. This is too large compared to the sample size. The second reason is that through later cumulative logit models analysis and the quasi-likelihood method analysis, we found that the interaction terms are not significant. Therefore, the full ME model we considered includes the gating network model

$$\log \frac{\pi_{ic}}{\pi_{i0}} = \gamma_{c0} + \gamma_{c1}\text{ICS} + \gamma_{c2}\text{CRO} + \gamma_{c3}\text{Assess} + \gamma_{c4}\text{Age}, \quad c = 1, \ldots, C.$$

and the expert network models, where the mean of the c^{th} simplex distribution is modeled by a logit model,

$$\text{logit}(\mu_c) = \beta_{c0} + \beta_{c1}\text{ICS} + \beta_{c2}\text{CRO} + \beta_{c3}\text{Assess} + \beta_{c4}\text{Age}, \quad c = 1, \ldots, C - 1.$$

We fitted the full ME models with $C = 2, 3, 4$, and used the model selection criteria AIC and AIC$_c$ to select the number of components. Table 3–14 gives the results when adding .01 to the number of 0% compliance to SRT. The AIC$_c$ value for $C = 3$ is the smallest one among these three models. The AIC value for $C = 3$ is close to the AIC value for $C = 4$. Since the sample size is relatively small compared to the number of parameter ($n/d < 40$), using the AIC$_c$ criterion is more appropriate. We chose $C = 3$ for this data set.

Table 3–14: The criterion of selecting C

C	$\ell(\widehat{\psi})$	No. of Para. d	AIC	AIC$_c$
2	-55.010	15	140.020	146.506
3	-17.412	25	84.824	105.137
4	-5.943	35	81.886	128.553

For fixed $C = 3$, we performed the likelihood-ratio tests for nested models. The final ME model includes

$$\log \frac{\pi_{ijc}}{\pi_{ij0}} = \gamma_{c0} + \gamma_{c1}\text{ICS} + \gamma_{c2}\text{CRO} + \gamma_{c3}\text{Assess}, \quad c = 1, 2, 3, \tag{3.43}$$

as the weights model,

$$\text{logit}(\mu_{i1}) = \beta_{10} + \beta_{11}\text{ICS} + \beta_{12}\text{CRO} + \beta_{13}\text{Assess} + \beta_{14}\text{Age}, \tag{3.44}$$

as the model for the mean of the first simplex distribution component, and

$$\text{logit}(\mu_{i2}) = \beta_{20}, \tag{3.45}$$

as the intercept model for the mean of the second simplex distribution component. Table 3–15 gives the ML parameter estimates after adding .01 for the number of 0% compliance to SRT.

The estimated parameters $\hat{\gamma}_{c3} > 0$ ($c = 1, 2, 3$) reveal that the higher the assessments, the larger the weights compared to the weight at mass point 0. For patients in the first distribution component, $\hat{\beta}_{11} = 2.831$ (S.E. = 0.156) suggests that ICS patients had higher compliances than SRT patients, $\hat{\beta}_{12} = -2.130$ (S.E. = 0.411) suggests us that CRO patients had lower compliances than SRT patients, and $\hat{\beta}_{11} - \hat{\beta}_{12} = 4.961$ (S.E. = 0.391) suggests that ICS patients had higher compliances than CRO patients, . For patients coming from the second simplex distribution component, medicines and doctor's assessment had no significant effects on patients' compliances.

Table 3–15: Parameter estimation of the ME model (C=3)

Parameter	ML estimate	S.E.
$\gamma_{10} - \gamma_{30}$ (Intercept)	6.119	1.829
$\gamma_{20} - \gamma_{30}$ (Intercept)	3.405	1.526
$\gamma_{11} - \gamma_{31}$ (ICS)	-3.205	1.006
$\gamma_{21} - \gamma_{31}$ (ICS)	-0.403	0.685
$\gamma_{12} - \gamma_{32}$ (CRO)	-0.851	1.096
$\gamma_{22} - \gamma_{32}$ (CRO)	-0.478	0.960
$\gamma_{11} - \gamma_{12}$ (ICS-CRO)	-0.145	1.649
$\gamma_{21} - \gamma_{22}$ (ICS-CRO)	2.284	1.709
$\gamma_{31} - \gamma_{32}$ (ICS-CRO)	2.210	1.759
γ_{13} (Assess)	1.522	1.200
γ_{23} (Assess)	2.533	1.182
γ_{33} (Assess)	3.422	1.256
β_{10} (Intercept)	-3.434	0.454
β_{11} (ICS-SRT)	2.831	0.156
β_{12} (CRO-SRT)	-2.130	0.411
$\beta_{11} - \beta_{12}$ (ICS-CRO)	4.961	0.391
β_{13} (Assess)	1.657	0.164
β_{14} (Age)	0.229	0.019
β_{20} (Intercept)	0.055	0.130
$\ell(\widehat{\psi})$	21.617	

Table 3–16: The estimated mean compliances conditional on assessment

Mean	Assess=1	Assess=2	Assess=3
ICS (S.E.)	0.486 (0.086)	0.638 (0.048)	0.722 (0.054)
CRO (S.E.)	0.085 (0.079)	0.369 (0.105)	0.689 (0.081)
SRT (S.E.)	0.397 (0.056)	0.693 (0.032)	0.760 (0.050)

Table 3–17: The estimated compliance differences conditional on assessment

Difference	Assess=1	Assess=2	Assess=3
SRT − ICS (S.E.)	-0.090 (.103)	0.055 (0.052)	0.038 (0.069)
SRT − CRO (S.E.)	0.312 (.089)	0.324 (0.110)	0.071 (0.093)
ICS − CRO (S.E.)	0.401 (.113)	0.269 (0.109)	0.034 (0.092)

Conditional on the assessment variable, and using the mean ages of the patients, we calculated mean compliances and the pairwise compliance differences for the three medicines.

$$\widehat{\mathrm{E}}(Y_s|\text{Assess} = q) = \sum_{c=1}^{C} \hat{\pi}_{sc}\hat{\mu}_{sc} \quad q = 1, 2, 3,$$

and

$$\hat{D}_{st}(\text{Assess} = q) = \widehat{\mathrm{E}}(Y_s|\text{Assess} = q) - \widehat{\mathrm{E}}(Y_t|\text{Assess} = q),$$

where s and t can be any two of the three medicines. Table 3–16 gives the estimated mean compliances of the three medicine conditional on the assessment variable. Table 3–17 gives the estimated compliance differences of the three medicine conditional on the assessment variable. The compliance to ICS and the compliance to SRT do not have a significant difference at all three assessment levels. At assessment level 1 and level 2, the mean compliances to ICS and SRT are significantly larger than the mean compliance to CRO. From Section 3.4.1, we know that the overall mean compliances of the three medicines did not have a significant difference. However, after we divided the patients into three levels according to the doctor's assessment, within level 1 and level 2, the differences between CRO and SRT , CRO and ICS are significant. Figure 3–3 shows the estimated compliances conditional on the doctor's assessment.

3.4.2.2 Cumulative logit models

We used the ungrouped data and the four-category data and the five-category data as described in Section 3.4.1 to fit the cumulative logit models. In all different settings, the same final model was chosen, which is

$$\text{logit}[P(Y_{i,g} \leq k)] = \theta_k - (\beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess}), \quad k = 1, 2, \ldots, K - 1.$$

Table 3–18 shows the model fitting results. Since the likelihood-ratio statistics are similar to the score statistics, the table only gives the score statistics.

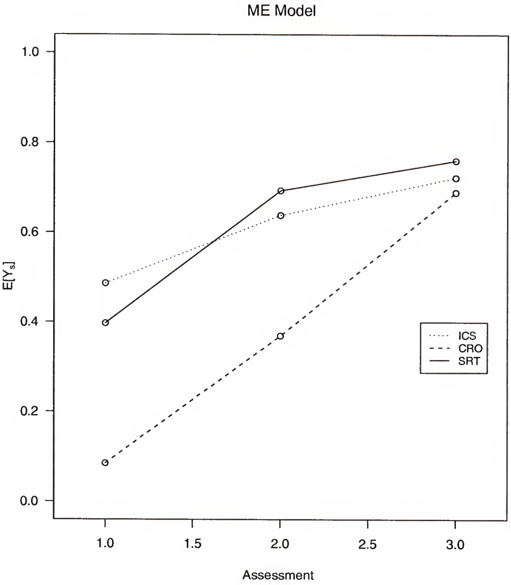


Figure 3–3: The estimated mean compliances for the three medicines conditional on the doctor's assessment based on the final ME model

In the final models, the score tests for the proportional odds assumption are not significant. Therefore, standard cumulative logit models are suitable to be used in fitting the grouped data sets. The standard errors of the estimated parameters from the three models are close. The Wald tests for $\hat{\beta}_1 = 0$ are not significant for all three models. This implies that the compliance to ICS was not significantly different from the compliance to SRT. The Wald tests for $\hat{\beta}_2 = 0$, $\hat{\beta}_3 = 0$, $\hat{\beta}_1 - \hat{\beta}_2 = 0$ are significant at approximately 0.05 significance level. This suggests that the compliance to CRO was significantly lower than the compliances to SRT and ICS conditional on doctor's assessment. The doctor's assessment can be used to predict the patient compliance. For example, for the four-category responses, $\hat{\beta}_2 = -1.455$ with a standard error of 0.643, $\hat{\beta}_1 - \hat{\beta}_2 = 1.349$ with a standard error of 0.627, and $\hat{\beta}_3 = 1.114$ with a standard error of 0.267 . The estimated odds that a patient's compliance to medicine SRT falls below any fixed category are $\exp(-1.455) = 0.233$ times the estimated odds of his compliance to CRO conditional on doctor's assessment. The estimated odds that a patient's compliance to medicine CRO falls below any fixed category are $\exp(1.349) = 3.854$ times the estimated odds of his compliance to ICS conditional on doctor's assessment. Conditional on a given medicine, the estimated odds that a patient's compliance falls below any fixed category decrease by $\exp(1.114) = 3.047$ if the doctor's assessment increases by 1.

3.4.2.3 Quasi-likelihood method

We used the quasi-likelihood method to fit the compliance data with other covariates. We again considered the two variance functions: $V_1(\mu_i) = \mu_i(1 - \mu_i)$ and $V_2(\mu_i) = [\mu_i(1 - \mu_i)]^2$. The final model is

$$\text{logit}(\mu_i) = \beta_0 + \beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess}.$$

Table 3–18: Parameter estimation for the cumulative logit models

Parameter	Ungrouped Estimate (S.E.)	4 categories Estimate (S.E.)	5 categories Estimate (S.E.)
β_1 (ICS-SRT)	0.069 (0.410)	-0.106 (0.452)	-0.246 (0.452)
β_2 (CRO-SRT)	-1.155 (0.597)	-1.455 (0.643)	-1.674 (0.644)
$\beta_1 - \beta_2$ (ICS-CRO)	1.225 (0.587)	1.349 (0.627)	1.428 (0.623)
β_3 (Assess)	1.193 (0.256)	1.114 (0.267)	1.266 (0.272)
T_1 (p-value)	—	4.666 (0.587) (df = 6)	9.741 (0.372) (df = 9)
T_2 (p-value)	19.761 (0.002)	18.418 (0.0004)	21.991 (<0.0001)

Note: T_1 = score test statistic for the proportional odds assumption;
T_2 = score test statistic for testing $\boldsymbol{\beta} = 0$.

Table 3–19: Parameter estimation for the quasi-likelihood methods

Parameter	$V_1(\mu_i) = \mu_i(1-\mu_i)$ Estimate (S.E.)	$V_2(\mu_i) = [\mu_i(1-\mu_i)]^2$ Estimate (S.E.)
β_0 (Intercept)	-1.079 (0.498)	-1.096 (0.518)
β_1 (ICS-SRT)	-0.180 (0.333)	-0.148 (0.323)
β_2 (CRO-SRT)	-0.996 (0.463)	-0.980 (0.476)
β_3 (Assess)	-0.808 (0.185)	-0.799 (0.186)
$\beta_1 - \beta_2$ (ICS-CRO)	0.817 (0.449)	0.833 (0.467)
ϕ	0.642	1.410
$T(\theta)$ (p-value)	1.880 (.170)	1.738 (0.187)

Note: $T(\theta)$ is the quasi-score statistic for testing the lack-of fit for the logit link function.

The model fitting results are given in Table 3–19. The two variance functions give the same final models and their parameter estimators and their standard errors are very close. For the family of link functions

$$g^{-1}(\boldsymbol{x}'\boldsymbol{\beta};\theta) = 1 - [1+\theta\exp(\boldsymbol{x}'\boldsymbol{\beta})]^{-1/\theta}, \qquad \theta \geq 0,$$

$\theta = 1$ reduces to the logit models we used. As shown in Table 3–19, the score tests for testing H_0: $\theta = 1$ suggest that the logit link is adequate.

Since the examination of the Pearson residuals can reveal the overall fit of the proposed model, we plotted the Pearson residuals against the linear predictor

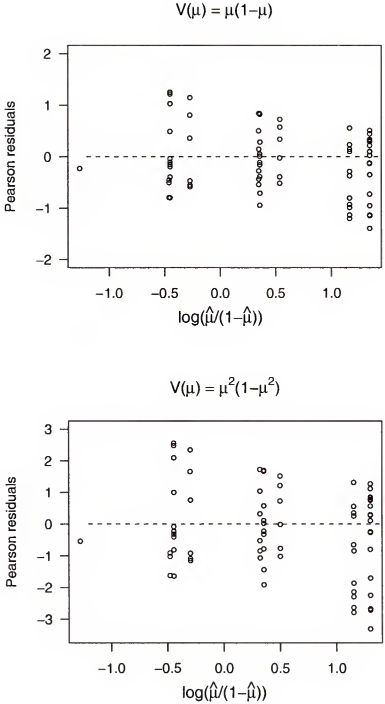
$\log(\hat{\mu}/(1-\hat{\mu}))$ for both variance functions (Figure 3–4). All the residuals for $V_1(\mu_i) = \mu_i(1-\mu_i)$ are between (-2, 2). Some of the residuals for $V_2(\mu_i) = [\mu_i(1-\mu_i)]^2$ are above 2 and below -2. Therefore, the quasi-likelihood method with variance function $\mu(1-\mu)$ seems to have a better fit. We used the estimated parameters for this variance function to estimate the mean compliances for the three medications conditional on the doctor's assessment and plotted Figure 3–5. Figure 3–5 strongly supports our conclusion that conditional on the doctor's assessment, the compliances of ICS patients and SRT patients had little difference, while the compliances of CRO patients were significantly lower than those of ICS patients and SRT patients.

We also used McCullagh (1983) test to compare the final model with the model without parameter β_1. For $V_1(\mu_i) = \mu_i(1-\mu_i)$, $2(Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})) = 0.24$. For $V_2(\mu_i) = [\mu_i(1-\mu_i)]^2$, $2(Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y})) = 0.82$. Therefore, the compliance of ICS patients was not significantly different from the compliance of SRT patients. The conclusion are the same as the ones using the cumulative logit model.

3.4.3 Summary

For analyzing the compliance data with three asthma medications, we proposed a ME model. It gave us great flexibility to handle this complex compliance data set. In addition, we considered two single-model methods to simplify the complex inference procedures in the ME models. These two single-model methods are the cumulative logit model and the quasi-likelihood method. We compared the three asthma medications with and without considering other covariates. Conclusions drawn from the three different analysis methods are consistent.

For this data set, we concluded that conditional on the doctor's assessment, the mean compliance to ICS was not significantly different from the mean compliance to SRT, and the mean compliance to CRO was significantly lower than the mean compliance to SRT and ICS. However, the ME final model also reveals



V(μ) = μ(1−μ)

V(μ) = μ²(1−μ²)

Figure 3-4: Pearson residuals plotted against the linear predictor $\log\left(\hat{\mu}/(1-\hat{\mu})\right)$

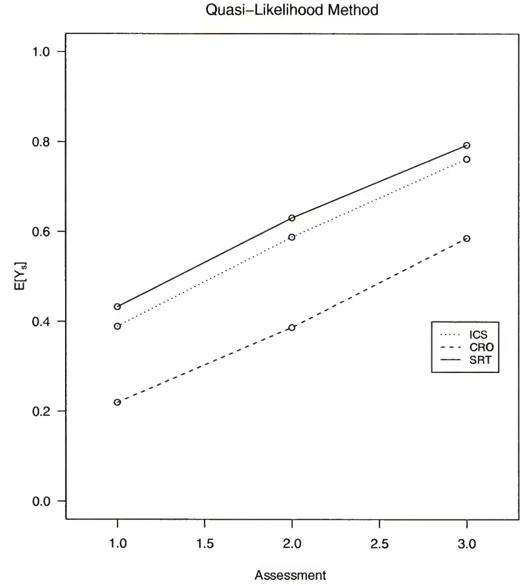


Figure 3–5: The estimated mean compliances for the three medicines conditional on the doctor's assessment based on the quasi-likelihood method with $V(\mu_i) = \mu_i(1 - \mu_i)$

that covariate effects are different in different sub-populations. For a group of lower compliance patients, the compliance to ICS was significant higher than the compliance to SRT ($\beta_{11} = 2.831$ with S.E. = 0.156); the compliance to CRO was significant lower than the compliance to SRT ($\beta_{12} = -2.130$ with S.E. = 0.411); the higher the doctor's assessment, the higher the compliance was ($\beta_{13} = 1.657$ with S.E. = 0.164); and older patients tended to have higher compliances than younger patients ($\beta_{13} = 0.229$ with S.E. = 0.019). But for a group of high compliance patients, the compliances were not significantly different with respect to the three medications, the doctor's assessment, and the patient's age. Although the single-model approaches are easier to apply, they don't have the ability to reveal the facts above.

Therefore, when one wants to compare the overall mean compliances of several groups, the single-model approaches are preferred. They are easier to fit and easier to use in making comparisons. When one wants to study the covariate effects on the compliances of the underlying sub-populations, the ME model is preferred.

CHAPTER 4
MODELING REPEATED MEASURES OF COMPLIANCE DATA

In medical compliance studies, researchers sometimes give the patients a combination of medicines to take. Each medicine has its own compliance rate. In other studies, compliances are observed for each subject repeatedly at various times. These repeated compliance responses are likely to be correlated. In this chapter, we extend the methods we proposed in Chapter 3 to analyze repeated measures of compliance data.

In the analysis of repeated measures of non-normal responses, two types of model approaches are used in different situations. The first type is the subject-specific model, which models the heterogeneity across subjects explicitly. The random effects models we discussed in Chapter 2 are such models. The other type is the population-averaged model, which models responses averaged over subjects without accounting for the heterogeneity. Marginal models belong to the latter. Agresti (2002 pp. 500-502) commented on the relationship between subject-specific models and population-averaged models. He noted that both types of models are useful and the choice of which type to use depends on the application. If one is interested in comparing the compliances between different groups, the population-averaged models are useful. If one is interested in estimating the expected changes in a subject's compliance when he switches to a different medicine, the subject-specific models are prefered.

In this chapter we consider both types of models in a repeated measures of compliance data study. Section 4.1 introduces a random effects ME model and uses a non-parametric approach in model fitting. Section 4.2 extends the random effects cumulative logit model we discussed in Chapter 2 to the random effects scaled cumulative logit model. Section 4.3 and 4.4 describe marginal models in the analysis of repeated measures of compliance data using the generalized estimating equation (GEE) approach. The review of the general GEE method and the GEE approach for grouped compliance data are given in Section 4.3. Section 4.4 discusses the GEE approach for the simplex model and proposes mixtures of marginal models. The last section (Section 4.5) uses two "real-life" data sets to illustrate the methods we propose.

4.1 Mixtures of Experts Model with Random Effects

4.1.1 Model Specification

Let us denote y_{ij} as the proportion of compliance at time j ($j = 1, \ldots, t_i$) for subject (or cluster) i ($i = 1, \ldots, n$). Since the two-part model is a special case of the ME model, we do not discuss the two-part model with random effects. We extend the ME model to correlated compliance data. Assume $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$ to be random effects designed to account for within-subject correlation, where $\boldsymbol{b}_{1i}$ are the random effects for the gating (weight) network model and $\boldsymbol{b}_{2i}$ are the random effects for the expert (component) network model. Let $\boldsymbol{x}_{ij}$ be a p-dimensional covariate vector pertaining to the fixed effects and $\boldsymbol{z}_{ij}$ be a q-dimensional covariate vector pertaining to the random effects for subject i at time j. As in the previous chapter, we assume that the number of expert networks is C.

Conditional on $\boldsymbol{b}_{1i}$, the gating network model is a multinomial logistic random effects model (Fahrmeir and Tutz, 2001)

$$\pi_{ijc} = \frac{\exp(\epsilon_{ijc})}{1 + \sum_{l=1}^{C} \exp(\epsilon_{ijl})},$$

with

$$\epsilon_{ijc} = \boldsymbol{x}_{ij}'\gamma_c + \boldsymbol{z}_{ij}'\boldsymbol{b}_{1ic}, \qquad c = 1, \ldots, C, \tag{4.1}$$

where $\boldsymbol{b}_{1i} = (\boldsymbol{b}'_{1i1}, \ldots, \boldsymbol{b}'_{1iC})'$ are independent from one subject to another following an unknown distribution with mean $\mathrm{E}(\boldsymbol{b}_{1i}) = \mathbf{0}$ and covariance matrix $\mathrm{cov}(\boldsymbol{b}_{1i}) = \boldsymbol{\Sigma}_{11}$.

Conditional on $\boldsymbol{b}_{2ic}$, the means of the simplex distributions are modeled by

$$\mathrm{logit}(\mu_{ijc}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_c + \boldsymbol{z}'_{ij}\boldsymbol{b}_{2ic}, \quad c = 1, \ldots, C-1. \tag{4.2}$$

We assume $\boldsymbol{b}_{2i} = (\boldsymbol{b}'_{2i1}, \ldots, \boldsymbol{b}'_{2i,C-1})'$ are independent from one subject to another following an unknown distribution with mean $\mathrm{E}(\boldsymbol{b}_{2i}) = \mathbf{0}$ and covariance matrix $\mathrm{cov}(\boldsymbol{b}_{2i}) = \boldsymbol{\Sigma}_{22}$. The random effects from the gating network model and the expert network model are possibly correlated. We assume that $\mathrm{cov}(\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i}) = \boldsymbol{\Sigma}_{12}$. In practice, the simple random intercept form of models are often adequate, in which $\boldsymbol{b}_{1i} = (b_{1i1}, \ldots, b_{1iC})'$, $\boldsymbol{b}_{2i} = (b_{2i1}, \ldots, b_{2i,C-1})'$ and $\boldsymbol{z}_{ij} = 1$.

Let $\boldsymbol{\psi}$ represent the unknown parameters, $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$. The marginal log-likelihood for the mixed effects ME model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\psi}),$$

with

$$\begin{aligned} L_i(\boldsymbol{\psi}) &= \int [\prod_{j=1}^{t_i} f(y_{ij}|\boldsymbol{b}_i)]\phi(\boldsymbol{b}_i)d\boldsymbol{b}_i \\ &= \int [\prod_{j=1}^{t_i}\sum_{c=0}^{C} \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{b}_{1i}) f_c(y_{ij}; \boldsymbol{\beta}_c|\boldsymbol{b}_{2ic})]\phi(\boldsymbol{b}_i)d\boldsymbol{b}_i, \end{aligned}$$

and $\phi()$ denotes the unknown density function for the random effects. We define

$$f_0(y_{ij}; \boldsymbol{\beta}_0|\boldsymbol{b}_{2i0}) = I(y_{ij} = 0),$$

$$f_C(y_{ij}; \boldsymbol{\beta}_C|\boldsymbol{b}_{2iC}) = I(y_{ij} = 1),$$

and

$$f_c(y_{ij};\boldsymbol{\beta}_c, |\boldsymbol{b}_{2ic}) = g(y_{ij};\boldsymbol{\beta}_c|\boldsymbol{b}_{2ic})I(0 < y_{ij} < 1), \quad c = 1, \ldots, C-1,$$

where $g()$ is the density function of the simplex distribution.

4.1.2 Nonparametric ML Model Fitting

Most random effects models assume that the random effects have a normal distribution. Considering the complexity of the ME model, if we assume that the random effects follow a multivariate normal distribution, the model fitting would be difficult. The ME model naturally accommodates the convenient EM algorithm. In Chapter 2, the nonparametric ML method of fitting a random effects model also applied the EM algorithm. Therefore, we now apply the NPML approach to fit the random intercept form of the mixed effects ME model.

We assume that $\phi()$ is an unknown discrete distribution with K mass points $\boldsymbol{m} = (\boldsymbol{m}_1', \ldots, \boldsymbol{m}_K')'$ and corresponding probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, where

$$\boldsymbol{m}_k = (\boldsymbol{m}_{1k}', \boldsymbol{m}_{2k}')',$$

$$\boldsymbol{m}_{1k} = (m_{1k1}, \ldots, m_{1kC})',$$

and

$$\boldsymbol{m}_{2k} = (m_{2k1}, \ldots, m_{2k,C-1})', \quad k = 1, \ldots, K.$$

Denote $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{m}', \boldsymbol{\pi}')$. The log-likelihood function is

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \Big[\prod_{j=1}^{t_i} f(y_{ij};\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{m}_k)\Big]$$

with

$$f(y_{ij};\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{m}_k) = \sum_{c=0}^{C} \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{m}_{1k}) f_c(y_{ij};\boldsymbol{\beta}_c|m_{2kc}).$$

For mass point k, we denote $\pi_{ijc|k} = \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{m}_{1k})$ and $\mu_{ijc|k} = \mu_{ijc}(\boldsymbol{\beta}_c|m_{2kc})$. The gating network is modeled by

$$\log \frac{\pi_{ijc|k}}{\pi_{ij0|k}} = \boldsymbol{x}_{ij}'\gamma_c + m_{1kc}, \qquad c = 1, \ldots, C, \tag{4.3}$$

and the expert network is modeled by

$$\operatorname{logit}(\mu_{ijc|k}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_c + m_{2kc}, \quad c = 1, \ldots, C-1. \tag{4.4}$$

Similar to the fitting of the ME model, we consider using the EM algorithm to obtain the NPML estimate for the mixed effects ME model. Assume that d_{ik} is an indicator that represents whether $\boldsymbol{y}_i$ is based on the k^{th} mass point, and $\Pr(d_{ik} = 1) = \pi_k$. Thus, $\sum_{k=1}^K d_{ik} = 1$ and $(d_{ik}|\boldsymbol{\pi})$ are *i.i.d.* with multinomial distribution $\prod_{k=1}^K \pi_k^{d_{ik}}$.

Given the mass point is k, assume that $d_{ijc|k}$ is an indicator function of whether $y_{ij}(\boldsymbol{m}_k)$ is generated from the c^{th} latent group, $\Pr(d_{ijc|k} = 1) = \pi_{ijc|k}$ and $\sum_{c=0}^C d_{ijc|k} = 1$. In general, the latent group from which y_{ij} is generated is not related to the random effects. Therefore, we can assume that $\{d_{ik}\}$ and $\{d_{ijc|k}\}$ are independent. We treat $\{d_{ik}\}$ and $\{d_{ijc|k}\}$ as missing data. The log-likelihood for the complete data is

$$\ell_{(c)}(\boldsymbol{\psi}) = \sum_{i=1}^n \sum_{k=1}^K d_{ik} \left\{ \sum_{j=1}^{t_i} \sum_{c=0}^C d_{ijc|k} \big[\log f_c(y_{ij}; \boldsymbol{\beta}_c|m_{2kc}) + \log \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{m}_{1k}) \big] + \log \pi_k \right\}.$$

At the $(t+1)^{th}$ E-step, replace the missing data by their expectation values conditional on the observed data and the parameter values at the t^{th} step.

$$\begin{aligned} \mathrm{E}[\ell_{(c)}(\boldsymbol{\psi}^{(t+1)}|\boldsymbol{y}, \boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^n \sum_{k=1}^K \sum_{j=1}^{t_i} \sum_{c=0}^C \mathrm{E}[d_{ik} d_{ijc|k}|\boldsymbol{y}, \boldsymbol{\psi}^{(t)}] \big[\log f_c(y_{ij}; \boldsymbol{\beta}_c|m_{2kc}) \\ &\quad + \log \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{m}_{1k}) \big] + \sum_{i=1}^n \sum_{k=1}^K \mathrm{E}[d_{ik}|\boldsymbol{y}, \boldsymbol{\psi}^{(t)}] \log \pi_k. \end{aligned}$$

Since $\{d_{ik}\}$ and $\{d_{ijc|k}\}$ are independent,

$$\mathrm{E}[d_{ik}d_{ijc|k}|\boldsymbol{y},\boldsymbol{\psi}^{(t)}] = \mathrm{E}[d_{ik}|\boldsymbol{y},\boldsymbol{\psi}^{(t)}]\mathrm{E}[d_{ijc|k}|\boldsymbol{y},\boldsymbol{\psi}^{(t)}] = w_{ik}^{(t)}w_{ijc|k}^{(t)},$$

where

$$w_{ijc|k}^{(t)} = \frac{\pi_{ijc}(\boldsymbol{\gamma}^{(t)}|\boldsymbol{m}_{1k}^{(t)})f_c(y_{ij};\boldsymbol{\beta}_c^{(t)}|m_{2kc}^{(t)})}{\sum_{l=0}^{C}\pi_{ijl}(\boldsymbol{\gamma}^{(t)}|\boldsymbol{m}_{1k}^{(t)})f_s(y_{ij};\boldsymbol{\beta}_l^{(t)}|m_{2kl}^{(t)})}, \quad c = 0,\ldots,C, \tag{4.5}$$

and

$$w_{ik}^{(t)} = \frac{\pi_k^{(t)}\left[\prod_{j=1}^{t_i} f(y_{ij};\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}|\boldsymbol{m}_k^{(t)})\right]}{\sum_{l=1}^{K}\pi_l^{(t)}\left[\prod_{j=1}^{t_i} f(y_{ij};\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}|\boldsymbol{m}_l^{(t)})\right]}, \quad k = 1,\ldots,K. \tag{4.6}$$

Thus, we can write the expected complete log-likelihood in closed form

$$\begin{aligned}
\mathrm{E}[\ell_{(c)}(\boldsymbol{\psi}^{(t+1)}|\boldsymbol{y},\boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc|k}^{(t)} \log f_c(y_{ij};\boldsymbol{\beta}_c|m_{2kc}) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc|k}^{(t)} \log \pi_{ijc}(\boldsymbol{\gamma}|\boldsymbol{m}_{1k}) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k. \\
&= h_1(\boldsymbol{\gamma},\boldsymbol{m}_1) + \sum_{c=1}^{C-1} h_2(\boldsymbol{\beta}_c,\boldsymbol{m}_{2c}) + h_3(\boldsymbol{\pi}),
\end{aligned}$$

where

$$h_1(\boldsymbol{\gamma},\boldsymbol{m}_1) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ik}^{(t)} w_{ijc|k}^{(t)} \log \pi_{ijc}(\boldsymbol{\gamma}_c|m_{1kc}), \tag{4.7}$$

$$h_2(\boldsymbol{\beta}_c,\boldsymbol{m}_{2c}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)} \log f_c(y_{ij};\boldsymbol{\beta}_c|m_{2kc}), \quad c = 1,\ldots,C-1, \tag{4.8}$$

and

$$h_3(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k. \tag{4.9}$$

At the $(t+1)^{th}$ M-Step, the expected complete log-likelihood is maximized with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(t+1)}$. Since $(\boldsymbol{\gamma},\boldsymbol{m}_1)$, $(\boldsymbol{\beta}_c,\boldsymbol{m}_{2c})$ and $\boldsymbol{\pi}$ are in three separate terms, the M-step reduces to the following separate maximization problems:

$$(\boldsymbol{\gamma}^{(t+1)},\boldsymbol{m}_1^{(t+1)})' = \boldsymbol{\psi}_1^{(t+1)} = \arg\max_{\boldsymbol{\psi}_1} h_1(\boldsymbol{\gamma},\boldsymbol{m}_1), \tag{4.10}$$

$$(\boldsymbol{\beta}_c^{(t+1)}, \boldsymbol{m}_{2c}^{(t+1)})' = \boldsymbol{\psi}_{2c}^{(t+1)} = \arg\max\nolimits_{\boldsymbol{\psi}_{2c}} h_2(\boldsymbol{\beta}_c, \boldsymbol{m}_{2c}), \quad c = 1, \dots, C-1, \tag{4.11}$$

$$\boldsymbol{\pi}^{(t+1)} = \arg\max\nolimits_{\boldsymbol{\pi}} h_3(\boldsymbol{\pi}). \tag{4.12}$$

When maximizing with respect to $\boldsymbol{\pi}$, we need to take the constraint $\sum_{k=1}^{K} \pi_k = 1$ into consideration. Solving the equations

$$\frac{\partial}{\partial \pi_k}[h_3(\boldsymbol{\pi}) - \lambda(\sum_{l=1}^{K} \pi_k - 1)] = \frac{1}{\pi_k} \sum_{i=1}^{n} w_{ik}^{(t)} - \lambda = 0$$

yields

$$\pi_k^{(t+1)} = \sum_{i=1}^{n} w_{ik}^{(t)} / n. \tag{4.13}$$

The maximization problem (4.10) with respect to $(\boldsymbol{\gamma}, \boldsymbol{m}_1)$ is a weighted version of the multinomial logit model ML estimators we discussed in the previous chapter. We re-define $\tilde{\boldsymbol{m}}_1 = (\tilde{\boldsymbol{m}}'_{11}, \dots,\ \boldsymbol{m}'_{1C})'$, where $\tilde{\boldsymbol{m}}_{1c} = (m_{11c}, \dots, m_{1Kc})'$. Let $\boldsymbol{Z}$ be a K-dimensional identity matrix. The multinomial logit model is

$$\log \frac{\pi_{ijc|k}}{\pi_{ij0|k}} = \boldsymbol{x}'_{ij}\gamma_c + \boldsymbol{z}'_k \tilde{\boldsymbol{m}}_{1c}, \qquad c = 1, \dots, C, \quad k = 1, \dots, K, \tag{4.14}$$

where $\boldsymbol{z}_k$ is the k^{th} column of matrix $\boldsymbol{Z}$. The ML estimates $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_C)'$ are the solutions to the score functions

$$\begin{aligned} \frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \gamma_c} &= \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{t_i} \sum_{l=0}^{C} w_{ik}^{(t)} w_{ijl|k}^{(t)} (\delta_{cl} - \pi_{ijc|k}) \boldsymbol{x}_{ij} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{t_i} w_{ik}^{(t)} (w_{ijc|k}^{(t)} - \pi_{ijc|k}) \boldsymbol{x}_{ij} = 0. \end{aligned}$$

The ML estimates of $\tilde{\boldsymbol{m}}_1$ are the solutions to the score functions

$$\frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_{1c}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{t_i} w_{ik}^{(t)} (w_{ijc|k}^{(t)} - \pi_{ijc|k}) \boldsymbol{z}_k = 0.$$

The maximization method is a hybrid Newton-Raphson method in which $\boldsymbol{\gamma}$ and $\tilde{\boldsymbol{m}}_1$ are alternately updated. The Newton-Raphson algorithm for updating $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} - H^{-1}(\boldsymbol{\gamma}^{(s)}) J(\boldsymbol{\gamma}^{(s)}), \tag{4.15}$$

where

$$J(\boldsymbol{\gamma}) = \frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \boldsymbol{\gamma}} = \Big(\frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \boldsymbol{\gamma}_1}, \ldots, \frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \boldsymbol{\gamma}_C}\Big)'$$

and

$$H(\boldsymbol{\gamma}) = \frac{\partial^2 h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \begin{bmatrix} H_{11}(\boldsymbol{\gamma}) & H_{12}(\boldsymbol{\gamma}) & \ldots & H_{1C}(\boldsymbol{\gamma}) \\ H_{21}(\boldsymbol{\gamma}) & H_{22}(\boldsymbol{\gamma}) & \ldots & H_{2C}(\boldsymbol{\gamma}) \\ \ldots & \ldots & \ldots & \ldots \\ H_{C1}(\boldsymbol{\gamma}) & H_{C2}(\boldsymbol{\gamma}) & \ldots & H_{CC}(\boldsymbol{\gamma}) \end{bmatrix},$$

in which

$$\begin{aligned} H_{cc'}(\boldsymbol{\gamma}) &= \frac{\partial^2 h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \boldsymbol{\gamma}_c \partial \boldsymbol{\gamma}'_{c'}} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}\pi_{ijc|k}(\delta_{cc'} - \pi_{ijc'|k})\boldsymbol{x}_{ij}\boldsymbol{x}'_{ij}, \\ & \qquad c, c' = 1, \ldots, C. \end{aligned}$$

The Newton-Raphson algorithm for updating $\tilde{\boldsymbol{m}}_1$ is

$$\tilde{\boldsymbol{m}}_1^{(s+1)} = \tilde{\boldsymbol{m}}_1^{(s)} - H^{-1}(\tilde{\boldsymbol{m}}_1^{(s)})J(\tilde{\boldsymbol{m}}_1^{(s)}), \tag{4.16}$$

where

$$J(\tilde{\boldsymbol{m}}_1) = \frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_1} = \Big(\frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_{11}}, \ldots, \frac{\partial h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_{1C}}\Big)'$$

and

$$H(\tilde{\boldsymbol{m}}_1) = \frac{\partial^2 h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_1 \partial \tilde{\boldsymbol{m}}'_1} = \begin{bmatrix} H_{11}(\tilde{\boldsymbol{m}}_1) & H_{12}(\tilde{\boldsymbol{m}}_1) & \ldots & H_{1C}(\tilde{\boldsymbol{m}}_1) \\ H_{21}(\tilde{\boldsymbol{m}}_1) & H_{22}(\tilde{\boldsymbol{m}}_1) & \ldots & H_{2C}(\tilde{\boldsymbol{m}}_1) \\ \ldots & \ldots & \ldots & \ldots \\ H_{C1}(\tilde{\boldsymbol{m}}_1) & H_{C2}(\tilde{\boldsymbol{m}}_1) & \ldots & H_{CC}(\tilde{\boldsymbol{m}}_1) \end{bmatrix},$$

in which

$$\begin{aligned} H_{cc'}(\tilde{\boldsymbol{m}}_1) &= \frac{\partial^2 h_1(\boldsymbol{\gamma}, \tilde{\boldsymbol{m}}_1)}{\partial \tilde{\boldsymbol{m}}_{1c} \partial \tilde{\boldsymbol{m}}'_{1c'}} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}\pi_{ijc|k}(\delta_{cc'} - \pi_{ijc'|k})\boldsymbol{z}_k\boldsymbol{z}'_k, \\ & \qquad c, c' = 1, \ldots, C. \end{aligned}$$

For $0 < y_{ij} < 1$, we need to solve the maximization problem (4.11). For each component, the solution is a weighted version of the ML estimator for the simplex model. Let us define $\tilde{\boldsymbol{m}}_{2c} = (m_{21c}, \ldots, m_{2Kc})'$. The model for the mean of the c^{th} component is

$$\operatorname{logit}(\mu_{ijc|k}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_c + \boldsymbol{z}_k'\tilde{\boldsymbol{m}}_{2c}, \quad c = 1, \ldots, C-1, \quad k = 1, \ldots, K. \tag{4.17}$$

As in the maximization method for problem (4.10), we also use a hybrid Newton-Raphson method to obtain the ML estimators for $\boldsymbol{\beta}$ and $\boldsymbol{m}_2$. The Newton-Raphson algorithm for updating $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}_c^{(s+1)} = \boldsymbol{\beta}_c^{(s)} - H^{-1}(\boldsymbol{\beta}_c^{(s)})J(\boldsymbol{\beta}_c^{(s)}), \quad c = 1, \ldots, C-1. \tag{4.18}$$

The Newton-Raphson algorithm for updating $\tilde{\boldsymbol{m}}_2$ is

$$\tilde{\boldsymbol{m}}_{2c}^{(s+1)} = \tilde{\boldsymbol{m}}_{2c}^{(s)} - H^{-1}(\tilde{\boldsymbol{m}}_{2c}^{(s)})J(\tilde{\boldsymbol{m}}_{2c}^{(s)}), \quad c = 1, \ldots, C-1. \tag{4.19}$$

In these formulas,

$$J(\boldsymbol{\theta}_c) = \frac{\partial h_2(\boldsymbol{\beta}_c, \tilde{\boldsymbol{m}}_{2c})}{\partial \boldsymbol{\theta}_c} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)} \left(-\frac{1}{2\sigma_c^2}\frac{\partial d(y_{ij}; \mu_{ijc|k})}{\partial \boldsymbol{\theta}_c} \right),$$

and

$$H(\boldsymbol{\theta}_c) = \frac{\partial^2 h_2(\boldsymbol{\beta}_c, \tilde{\boldsymbol{m}}_{2c})}{\partial \boldsymbol{\theta}_c \partial \boldsymbol{\theta}_c'} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)} \left(-\frac{1}{2\sigma_c^2}\frac{\partial^2 d(y_{ij}; \mu_{ijc|k})}{\partial \boldsymbol{\theta}_c \partial \boldsymbol{\theta}_c'} \right),$$

where $\boldsymbol{\theta}_c = \boldsymbol{\beta}_c$ or $\tilde{\boldsymbol{m}}_{2c}$.

Since

$$\frac{\partial h_2(\boldsymbol{\beta}_c, \tilde{\boldsymbol{m}}_{2c})}{\partial \sigma_c^2} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)} \left(-\frac{1}{2\sigma_c^2} + \frac{1}{2\sigma_c^4} d(y_{ij}; \mu_{ijc|k}) \right),$$

$$\sigma_c^{2(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)} d(y_{ij}; \mu_{ijc|k}^{(t+1)})}{\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{t_i} w_{ik}^{(t)} w_{ijc|k}^{(t)}}. \tag{4.20}$$

The EM algorithm is summarized as follows:

0. Specify initial values $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{m}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$. Specify tolerances values ϵ and ϵ_1.

1. Increase t by 1. (E-Step) For the i^{th} subject at time j, $i = 1, \ldots, n$, $j = 1, \ldots, t_i$, compute the weights $w_{ijc|k}$ $(c = 0, \ldots, C,\ k = 1, \ldots, K)$ using (4.5) and w_{ik} $(k = 1, \ldots, K)$ using (4.6).

2. (M-step) Solve the ML estimators for $\boldsymbol{\pi}$ using (4.13). Solve the ML estimators for $\boldsymbol{\psi}_1 = (\boldsymbol{\gamma}', \boldsymbol{m}_1')'$ in (4.10) using the convergence criterion

$$\frac{||\boldsymbol{\psi}_1^{(t)} - \boldsymbol{\psi}_1^{(t-1)}||}{||\boldsymbol{\psi}_1^{(t-1)}||} \leq \epsilon_1.$$

Solve the ML estimator for $\boldsymbol{\psi}_{2c} = (\boldsymbol{\beta}_c', \boldsymbol{m}_{2c}')'$ $(c = 1, \ldots, C-1)$ in (4.11) using the convergence criterion

$$\frac{||\boldsymbol{\psi}_{2c}^{(t)} - \boldsymbol{\psi}_{2c}^{(t-1)}||}{||\boldsymbol{\psi}_{2c}^{(t-1)}||} \leq \epsilon_1.$$

3. Iterate between 1 and 2 until the overall convergence criteria is satisfied, which is

$$|\ell(\boldsymbol{\psi}^{(t)}) - \ell(\boldsymbol{\psi}^{(t-1)})| \leq \epsilon,$$

where $\ell(\boldsymbol{\psi}^{(t)})$ is the observed log-likelihood function.

4.1.3 Standard Error Estimation

The calculation of the standard error estimates is more complicated than that of the previous chapter. We again use Louis' method (1982) to approximate the observed information matrix. Assume that $\boldsymbol{\psi}_* = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$. The complete log-likelihood function is

$$\begin{aligned}
\ell_{(c)}(\boldsymbol{\psi}_*) &= \sum_{i=1}^{n} \ell_{i(c)}(\boldsymbol{\psi}_*) = \sum_{i=1}^{n}\sum_{k=1}^{K} \ell_{ik(c)}(\boldsymbol{\psi}_*) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} d_{ik} \sum_{j=1}^{t_i} \ell_{ijk(c)}(\boldsymbol{\psi}_*) + \sum_{i=1}^{n}\sum_{k=1}^{K} d_{ik} \log \pi_k \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} d_{ik} \sum_{j=1}^{t_i}\sum_{l=0}^{C} d_{ijl|k} \ell_{ijl|k}(\boldsymbol{\psi}_*) + \sum_{i=1}^{n}\sum_{k=1}^{K} d_{ik} \log \pi_k .
\end{aligned}$$

where

$$\ell_{ijl|k}(\boldsymbol{\psi}_*) = \log f_l(y_{ij}|\boldsymbol{\beta}_l) + \log \pi_{ijl|k}(\boldsymbol{\gamma}).$$

The observed variance-covariance matrix is then estimated by

$$\mathrm{E}\left[-\frac{\partial^2 \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_* \partial \boldsymbol{\psi}_*'}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] - \mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right]. \tag{4.21}$$

The estimated information matrix is

$$\mathrm{E}\left[-\frac{\partial^2 \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_* \partial \boldsymbol{\psi}_*'}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] = \begin{bmatrix} -H(\hat{\boldsymbol{\gamma}}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{H}(\hat{\boldsymbol{\beta}}_1) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & -\mathbf{H}(\hat{\boldsymbol{\beta}}_{\mathbf{C-1}}) \end{bmatrix}, \tag{4.22}$$

where $H(\hat{\boldsymbol{\gamma}})$ and $H(\hat{\boldsymbol{\beta}}_c)$ $(c = 1, \dots, C-1)$ are the estimated second derivatives in the M-step.

The second term in (4.21) is

$$\begin{aligned}
& \mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] && (4.23) \\
= \ & \mathrm{E}\left[\left(\sum_{i=1}^{n}\frac{\partial \ell_{i(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\right)\left(\sum_{i=1}^{n}\frac{\partial \ell_{i(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\right)\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] && (4.24) \\
= \ & \sum_{i=1}^{n}\mathrm{E}\left[\left(\sum_{k=1}^{K}\frac{\partial \ell_{ik(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\right)\left(\sum_{k=1}^{K}\frac{\partial \ell_{ik(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\right)\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] && (4.25) \\
& + \sum_{i=1}^{n}\sum_{i\neq i'}\mathrm{E}\left[\frac{\partial \ell_{i(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right]\mathrm{E}\left[\frac{\partial \ell_{i'(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\bigg| \boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right]. && (4.26)
\end{aligned}$$

Since $d_{ik}d_{ik'} = d_{ik}$ only if $k = k'$, and $d_{ik}d_{ik'} = 0$ otherwise, equation (4.25) equals

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathrm{E}\Big[\sum_{k=1}^{K} d_{ik}\Big(\sum_{j=1}^{t_i} \frac{\partial \ell_{ijk(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\Big)\Big(\sum_{j'=1}^{t_i} \frac{\partial \ell_{ij'k(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\Big)\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \\
= &\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik} \sum_{j=1}^{t_i}\sum_{j\neq j'} \mathrm{E}\Big[\frac{\partial \ell_{ijk(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \mathrm{E}\Big[\frac{\partial \ell_{ij'k(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \\
&+\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik} \sum_{j=1}^{t_i} \mathrm{E}\Big[\Big(\sum_{l=0}^{C} d_{ijl|k} \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\Big)\Big(\sum_{l'=0}^{C} d_{ijl'|k} \frac{\partial \ell_{ijl'|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\Big)\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \\
= &\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik} \sum_{j=1}^{t_i}\sum_{j\neq j'} \mathrm{E}\Big[\frac{\partial \ell_{ijk(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \mathrm{E}\Big[\frac{\partial \ell_{ij'k(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\Big|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\Big] \\
&+\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{w}_{ik} \sum_{j=1}^{t_i}\sum_{l=0}^{C} \hat{w}_{ijl|k} \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*} \frac{\partial \ell_{ijl'|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}.
\end{aligned}
$$

The last step holds because $d_{ijl|k}d_{ijl'|k} = d_{ijl|k}$ only when $l = l'$, and $d_{ijl|k}d_{ijl'|k} = 0$ otherwise.

Let us define

$$
S_{ijl|k}(\boldsymbol{\psi}_*) = \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*} = \Big(\frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\gamma}_1}, \ldots, \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\gamma}_C}, \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\beta}_1}, \ldots, \frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\beta}_{C-1}}\Big)',
$$

where

$$
\frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\gamma}_c} = (\delta_{cl} - \pi_{ijc|k})\boldsymbol{x}_{ij}, \quad c = 1, \ldots, C
$$

and

$$
\frac{\partial \ell_{ijl|k}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\beta}_c} = \begin{cases} -\frac{1}{2\sigma_l^2}\frac{\partial d(y_{ij};\mu_{ijl|k})}{\partial \boldsymbol{\beta}_c} & \text{if } l = c \\ 0 & \text{if } l \neq c \end{cases} \qquad c = 1, \ldots, C-1.
$$

Therefore,

$$\mathrm{E}\left[\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*}\frac{\partial \ell_{(c)}(\boldsymbol{\psi}_*)}{\partial \boldsymbol{\psi}_*'}\middle|\boldsymbol{y}, \widehat{\boldsymbol{\psi}}\right] =$$

$$\sum_{i=1}^{n}\sum_{i\neq i'}\Big(\sum_{k=1}^{K}\sum_{j=1}^{t_i}\sum_{l=0}^{C}\hat{w}_{ik}\hat{w}_{ijl|k}S_{ijl|k}(\widehat{\boldsymbol{\psi}}_*)\Big)\Big(\sum_{k'=1}^{K}\sum_{j'=1}^{t_i}\sum_{l'=0}^{C}\hat{w}_{i'k'}\hat{w}_{i'j'l'|k'}S_{i'j'l'|k'}(\widehat{\boldsymbol{\psi}}_*)\Big)'$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}\sum_{j=1}^{t_i}\sum_{j\neq j'}\Big(\sum_{l=0}^{C}\hat{w}_{ijl|k}S_{ijl|k}(\widehat{\boldsymbol{\psi}}_*)\Big)\Big(\sum_{l'=0}^{C}\hat{w}_{ij'l'|k}S_{ij'l'|k}(\widehat{\boldsymbol{\psi}}_*)\Big)'$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{w}_{ik}\sum_{j=1}^{t_i}\sum_{l=0}^{C}\hat{w}_{ijl|k}S_{ijl|k}(\widehat{\boldsymbol{\psi}}_*)S'_{ijl|k}(\widehat{\boldsymbol{\psi}}_*).$$

Then, the standard error estimates can be obtained by inverting the estimated observed information matrix, and taking the square root of the diagonal elements.

4.1.4 Model Selection

First, we let $K = 1$, which results in a cross-sectional ME model. We use the two-step model selection method we proposed in the previous chapter to choose C and then determine the number of parameters. We calculate the maximum log-likelihood for $\sum_{i=1}^{n} t_i$ independent responses and define it as $\ell_1(\widehat{\boldsymbol{\psi}})$. For the chosen C and the chosen covariates, given a choice of the number K ($K > 2$), we calculate the maximized log-likelihood by

$$\ell_K(\widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\hat{\pi}_k[\prod_{j=1}^{t_i} f(y_{ij};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{m}}_k)], \tag{4.27}$$

where

$$f(y_{ij};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{m}}_k) = \sum_{c=0}^{C}\hat{\pi}_{ijc}(\hat{\boldsymbol{\gamma}}|\hat{\boldsymbol{m}}_{1k})f_c(y_{ij};\hat{\boldsymbol{\beta}}_c|\hat{m}_{2kc}).$$

We define the deviance difference, comparing this model to the simpler non-mixture model, by

$$\mathrm{dev}_K = 2[\ell_K(\widehat{\boldsymbol{\psi}}) - \ell_1(\widehat{\boldsymbol{\psi}})].$$

For the simpler non-mixture model, $\pi_2, \ldots, \pi_K$ are all zeros, which are at the boundary of the parameter space. The standard likelihood-ratio test cannot be applied here. We estimated the number of mass points K by starting with $K = 2$ and increasing K until the change in the deviance is small.

4.2 Scaled Cumulative Logit Models with Random Effects

We discussed cumulative logit models with random effects in Chapter 2. For scaled cumulative logit models with random effects, we group the proportions of compliances into K ordered groups and let $Y_{ij,g}$ be the grouped response variable. Conditional on the random effects $\boldsymbol{b}_i$, the model has the form

$$\operatorname{logit}[P(Y_{ij,g} \leq k)] = \eta_{ijk} = \frac{\theta_k - (\boldsymbol{x}'_{1ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i)}{\exp(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma})}, \quad k = 1, 2, \ldots, K-1, \qquad (4.28)$$

where $\boldsymbol{b}_i \sim N(0, \boldsymbol{\Sigma})$. The probability that $Y_{ij,g}$ takes on the value k is

$$\pi_{ij,k} = P(Y_{ij,g} = k) = \frac{1}{1 + \exp(-\eta_{ijk})} - \frac{1}{1 + \exp(-\eta_{ij,k-1})} \quad k = 1, 2, \ldots, K,$$

with $\eta_{ij0} = -\infty$.

The conditional probability for the response variable $Y_{ij,g}$ is

$$f(y_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{b}_i) = \prod_{k=1}^{K} \pi_{ij,k}^{d_{ijk}},$$

where $d_{ijk} = 1$ if $Y_{ij,g} = k$ and $d_{ijk} = 0$ otherwise. The marginal log-likelihood for this model is

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \int \prod_{j=1}^{t_i} [f(y_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{b}_i] \phi(\boldsymbol{b}_i) d\boldsymbol{b}_i.$$

As in Chapter 2, the ML estimate can be solved by the Fisher scoring method. The SAS procedure NLMIXED can be used to fit the model.

4.3 Marginal Models and GEE

In the previous two sections, we analyzed the repeated measures of the compliance data at the subject level. If inferences about the average compliance over all subjects are the study focus, marginal models are more appropriate to use. In the random effects models in the previous section, we modeled the covariate effects and within-subject association through a single model. In contrast with the random effects model, the marginal models model the covariate effects and the relationship between the observations on the same subject separately. Marginal models for non-normal longitudinal data were first proposed by Liang and Zeger (1986) and Zeger and Liang (1986). They introduced the generalized estimating equations (GEE) method. We first introduce the basic GEE method, which is the marginal model for the quasi-likelihood method we proposed in the previous chapter. To extend the cumulative logit model for the grouped compliance data to the marginal model setting, we then introduce the marginal model for repeated ordinal response data.

4.3.1 Generalized Estimating Equations

For subject i, let $\boldsymbol{Y} = (Y_{i1}, \ldots, Y_{it_i})'$ be the response vector and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{it_i})'$ be the $t_i \times p$ covariate matrix. Since the GEE method is an extension of the quasi-likelihood (QL) method, one does not need to specify the underlying distribution of the observations. As in the QL method, we assume that the marginal mean of Y_{ij} is

$$\mathrm{E}(Y_{ij}) = \mu_{ij} = g^{-1}(\boldsymbol{x}_{ij}'\boldsymbol{\beta}). \tag{4.29}$$

The marginal variance is a function of μ_{ij},

$$\mathrm{Var}(Y_{ij}) = \phi V(\mu_{ij}), \tag{4.30}$$

where ϕ is the dispersion parameter. In addition to the above two assumptions for the QL model, we also need to assume a working correlation matrix $R(\boldsymbol{\alpha})$ for $\boldsymbol{Y}_i$, where $\boldsymbol{\alpha}$ are the association parameters. The working covariance matrix for $\boldsymbol{Y}_i$ is specified by

$$\boldsymbol{\Sigma}_i = \phi \boldsymbol{V}_i^{1/2} R(\boldsymbol{\alpha}) \boldsymbol{V}_i^{1/2}, \tag{4.31}$$

where $\boldsymbol{V}_i = \mathrm{diag}[V(\mu_{i1}), \ldots, V(\mu_{it_i})]$. The working correlation matrix $R(\boldsymbol{\alpha})$ can be parameterized with an independence correlation structure, an exchangeable correlation structure, or an autoregressive correlation structure, among others. If

$R(\boldsymbol{\alpha})$ is the true correlation matrix for $\boldsymbol{Y}_i$, $\boldsymbol{\Sigma}_i = \text{cov}(\boldsymbol{Y}_i)$. In practice, we would not expect this, so we use a robust covariance matrix that empirically adjusts the working one.

Analogous to the quasi-score function for the QL model, the generalized estimating equations for $\boldsymbol{\beta}$ are

$$\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{4.32}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$.

Liang and Zeger (1986) suggested computing the GEE estimate $\hat{\boldsymbol{\beta}}$ by iterating between a modified Fisher scoring algorithm and estimation for $\boldsymbol{\alpha}$ and ϕ. First one needs to select a working correlation matrix. Given the current estimates of $\boldsymbol{\alpha}$ and ϕ, the approximate estimate for $\boldsymbol{\beta}$ is solved by iteration

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \big\{ \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \boldsymbol{D}_i' \big\}^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}). \tag{4.33}$$

Given the correct specification of the mean and some regularity conditions, the GEE estimate $\hat{\boldsymbol{\beta}}$ is consistent and asymptotic normal as $n \to \infty$. The robust covariance estimator of $\hat{\boldsymbol{\beta}}$ (Zeger and Liang 1986) is

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \hat{I}_0^{-1} \hat{I}_1 \hat{I}_0^{-1},$$

where $I_0 = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \boldsymbol{D}_i$, and $I_1 = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \text{cov}(\boldsymbol{Y}_i) \Sigma_i^{-1} \boldsymbol{D}_i$ with $\text{cov}(\boldsymbol{Y}_i) = (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$.

For the estimation of $\boldsymbol{\alpha}$ and ϕ, there are several possibilities (Prentice 1988, Liang and Zeger 1986). In our data analysis, we used the Liang and Zeger moment estimation method. For each iteration, we use the current estimation of $\hat{\boldsymbol{\beta}}$ to compute Pearson residuals

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{\mu}_{ij})}}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, t_i.$$

The dispersion parameter is estimated by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{t_i} \hat{r}_{ij}^2,$$

where $N = \sum_{i=1}^{n} t_i$ and the p is the number of estimated parameters. For a specific choice of $R(\boldsymbol{\alpha})$, one estimates $\boldsymbol{\alpha}$ correspondingly. When $R(\boldsymbol{\alpha}) = I$, one treats the responses as independent. For exchangeable correlation, $\text{corr}(Y_{ij},\, Y_{ij'}) = \alpha$ for all $j \neq j'$, α is estimated by

$$\hat{\alpha} = \frac{1}{\hat{\phi}\{\sum_{i=1}^{n} \frac{1}{2} t_i(t_i - 1) - p\}} \sum_{i=1}^{n} \sum_{j'>j} \hat{r}_{ij'} \hat{r}_{ij}.$$

If one specifies an unstructured correlation matrix, $\boldsymbol{\alpha}$ is estimated by

$$\hat{\alpha}_{jj'} = \frac{1}{\hat{\phi}(n-p)} \sum_{i=1}^{n} \hat{r}_{ij} \hat{r}_{ij'}, \quad j, j' = 1, \ldots, t_i.$$

The SAS procedure GENMOD can be used to obtain the GEE estimates.

4.3.2 GEE for Repeated Ordinal Response Data

For the cross-sectional data analysis, we proposed one method that groups the possible outcomes into K ordered groups and uses a cumulative logit model to analyze the ordinal grouped data. In the repeated ordinal data setting, we need to extend the GEE method for repeated measures. Lipsitz et al. (1994) generalized the GEE method to models for repeated ordinal response data. In our compliance data analysis, we used that approach.

Let $Y_{ijk,g} = 1$ if the grouped response for subject i at time j belongs to the k^{th} ordered group, and $Y_{ijk,g} = 0$ if otherwise. We form the $(K-1) \times 1$ vector $\boldsymbol{Y}_{ij,g} = (Y_{ij1,g}, \ldots, Y_{ijK-1,g})'$. Let $\pi_{ijk} = Pr(Y_{ijk,g} = 1)$. The response vector $\boldsymbol{Y}_{ij,g}$ follows a multinomial distribution $\sum_{k=1}^{K} \pi_{ijk}^{y_{ijk,g}}$. Therefore, the marginal mean of $\boldsymbol{Y}_{ij,g}$ is $\boldsymbol{\pi}_{ij}$, which can be modeled by a cumulative logit model,

$$\text{logit}\Big[\sum_{l=1}^{k} \pi_{ijl}\Big] = \theta_k - \boldsymbol{x}'_{ij}\boldsymbol{\beta}, \quad k = 1, 2, \ldots, K-1. \tag{4.34}$$

The marginal covariance matrix $V_{ij} = \mathrm{Var}(\boldsymbol{Y}_{ij,g}) = \mathrm{diag}[\boldsymbol{\pi}_{ij}] - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}'_{ij}$, where $\mathrm{diag}[\boldsymbol{\pi}_{ij}] = \mathrm{diag}[\pi_{ij1}, \ldots, \pi_{ij,K-1}]$.

Let us form the $t_i(K-1) \times 1$ vector $\boldsymbol{Y}_{i,g} = (\boldsymbol{Y}'_{i1,g}, \ldots, \boldsymbol{Y}'_{it_i,g})'$. The generalized estimating equations for the grouped ordinal response data are

$$\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}'_i \Sigma_i^{-1} (\boldsymbol{y}_{i,g} - \boldsymbol{\mu}_i), \tag{4.35}$$

where $\boldsymbol{\pi}_i = (\boldsymbol{\pi}'_{i1}, \ldots, \boldsymbol{\pi}'_{it_i})'$, $\boldsymbol{D}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i^{-1}$ is a $t_i(K-1) \times t_i(K-1)$ working covariance matrix for $\boldsymbol{y}_{i,g}$. The $(K-1) \times (K-1)$ diagonal blocks of $\boldsymbol{\Sigma}_i$ are V_{ij}, $j = 1, \ldots, t_i$. Let us define a $(K-1) \times (K-1)$ diagonal matrix $\boldsymbol{B}_{ij}$ by

$$\boldsymbol{B}_{ij} = \mathrm{diag}[\pi_{ij1}(1 - \pi_{ij1}), \ldots, \pi_{ij,K-1}(1 - \pi_{ij,K-1})].$$

and let $\boldsymbol{B}_i = \mathrm{diag}\{\boldsymbol{B}_{i1}, \ldots, \boldsymbol{B}_{it_i}\}$. The working covariance matrix is formed by

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i^{1/2} R_i(\boldsymbol{\alpha}) \boldsymbol{B}_i^{1/2},$$

where the working correlation matrix is

$$R_i(\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{B}_{i1}^{-1/2} V_{i1} \boldsymbol{B}_{i1}^{-1/2} & \boldsymbol{\rho}_{i12} & \cdots & \boldsymbol{\rho}_{i1t_i} \\ \boldsymbol{\rho}_{i21} & \boldsymbol{B}_{i2}^{-1/2} V_{i2} \boldsymbol{B}_{i2}^{-1/2} & \cdots & \boldsymbol{\rho}_{i2t_i} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ \boldsymbol{\rho}_{it_i1} & \boldsymbol{\rho}_{it_i2} & \cdots & \boldsymbol{B}_{it_i}^{-1/2} V_{it_i} \boldsymbol{B}_{it_i}^{-1/2} \end{bmatrix}.$$

with $\boldsymbol{\rho}_{ijj'}$ $(j \neq j')$ representing the working correlation matrix between $\boldsymbol{Y}_{ij,g}$ and $\boldsymbol{Y}_{ij',g}$. When $\boldsymbol{\rho}_{ijj'} = \boldsymbol{0}$ for all $j \neq j'$, $\boldsymbol{\Sigma}_i = \mathrm{diag}[V_{i1}, \ldots, V_{it_i}]$, one treats the responses as independent.

We extend the method of moments to estimate the correlation matrix $\boldsymbol{\alpha}$. The residual vector for subject i at time j is

$$\boldsymbol{r}_{ij} = \boldsymbol{B}_{ij}^{-1/2}(\boldsymbol{y}_{ij,g} - \boldsymbol{\pi}_{ij}).$$

Under the exchangeable correlation structure, $\rho_{ijj'} = \alpha$ for all i, j and j', α is estimated by

$$\hat{\alpha} = \frac{1}{\{\sum_{i=1}^{n} \frac{1}{2} t_i(t_i - 1)\} - p} \sum_{i=1}^{n} \sum_{j'>j} \hat{r}_{ij'} \hat{r}'_{ij}.$$

Under the unstructured correlation assumption, $\rho_{ijj'} = \alpha_{jj'}$ for all i, α is estimated by

$$\hat{\alpha}_{jj'} = \frac{1}{(n-p)} \sum_{i=1}^{n} \hat{r}_{ij} \hat{r}'_{ij'}, \quad j, j' = 1, \ldots, t_i.$$

The SAS procedure GENMOD gives only the estimates for an independence working correlation structure.

4.4 Mixtures of Marginal Models

In this section we extend the ME model to the marginal model setting. We use the simplex distributions as the components in the ME model. Since the simplex distribution does not belong to the exponential dispersion model family, the regular GEE method cannot be applied to the simplex distribution. We extend the GEE method to the simplex distribution first. Then we combine this extension of the GEE method with the ME model we proposed in the previous chapter to form the mixtures of marginal models. A generalization of the EM algorithm is introduced to fit the model.

4.4.1 Extension of GEE Method for Simplex Distributions

Recall from the previous chapter, the density function of the simplex distribution $S^{-}(\mu, \sigma^2)$ is defined as

$$f(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2} d(y; \mu)\},$$

for $0 < y < 1$, where

$$d(y; \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},$$

$0 < \mu < 1$, and $\sigma^2 > 0$. The mean of the response variable Y is $\mathrm{E}(Y) = \mu$.

The dispersion model introduced by Jørgensen (1997) is defined as

$$f(y;\mu,\sigma^2)=a(\sigma^2,y)\mathrm{exp}\bigg\{-\frac{1}{2\sigma^2}d(y;\mu)\bigg\}.$$

where $a()$ is a suitable function and $d(y;u)$ is a regular unit deviance. The unit variance function of the dispersion model is defined as (Jørgensen, 1997),

$$V(\mu)=\frac{2}{\frac{\partial^2 d}{\partial\mu^2}(\mu;\mu)}$$

The simplex distribution belongs to the dispersion models. The unit variance function for the simplex distribution is $V(\mu)=\mu^3(1-\mu)^3$.

The exponential family is a special family of the dispersion models. The density function of the exponential family has the form

$$f(y)=\exp\bigg\{-\frac{1}{2\sigma^2}\big(y\theta-a(\theta)+b(y)\big)\bigg\}.$$

The first two moments of Y are $\mathrm{E}(Y)=\mu=a'(\theta)$ and $\mathrm{Var}(Y)=\sigma^2V(\mu)=\sigma^2a''(\theta)$.

The GEE method proposed by Liang and Zeger (1986) are derived from the exponential family models. This method needs the assumption that the marginal variance of the response is a function of the the unit variance function (Jørgensen, 1997), i.e., $\mathrm{Var}(Y)=\phi V(\mu)$. But the variance of the simplex distribution is

$$\mathrm{Var}(Y)=\mu(1-\mu)-\frac{1}{\sqrt{2}\sigma}\exp\{\frac{1}{\sigma^2\mu^2(1-\mu)^2}\}\times\Gamma\{\frac{1}{2},\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\},$$

where $\Gamma(a,b)=\int_b^\infty x^{a-1}e^{-x}dx$ is an incomplete gamma function (Song 2000). Thus, the standard GEE method cannot be applied for the simplex distribution. Artes and Jørgensen (2000) extended the GEE method for the dispersion models. This method replaces the score function $(Y-\mu)/V(\mu)$ for the exponential family by the dispersion model score function $u(Y;\mu)$.

We apply the extension of the GEE method for dispersion models (Artes and Jørgensen, 2000) to the simplex distribution. Assume that $Y_{ij}\sim S^-(\mu_{ij},\sigma^2)$, a simplex distribution. The marginal mean is modeled by $\mu_{ij} = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$. We define $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{it_i})'$ to be the score vector, with

$$u_{ij} = -\frac{1}{2}d'(y_{ij};\mu_{ij}) = \frac{y_{ij}-\mu_{ij}}{\mu_{ij}(1-\mu_{ij})}\Big\{d(y_{ij};\mu_{ij}) + \frac{1}{\mu_{ij}^2(1-\mu_{ij})^2}\Big\}. \tag{4.36}$$

Song and Tan (2000) proved the following properties of the simplex distribution:

Let $Y \sim S^-(\mu, \sigma^2)$. Then

(1) $\mathrm{E}d(y;\mu) = \sigma^2$;

(2) $\mathrm{E}(y-\mu)d'(y;\mu) = -2\sigma^2$;

(3) $\mathrm{E}(y-\mu)d(y;\mu) = 0$;

(4)$(1/2)\mathrm{E}d''(y;\mu) = 3\sigma^2/[\mu(1-\mu)] + 1/[\mu^3(1-\mu)^3]$.

Following properties (3) and $\mathrm{E}(y_{ij}) = \mu_{ij}$, we can show that $\mathrm{E}(u_{ij}) = 0$. Since

$$\mathrm{E}\Big[\Big(\frac{\partial \log f(y_{ij};\mu_{ij},\sigma^2)}{\partial \mu_{ij}}\Big)^2\Big] = -\mathrm{E}\Big[\frac{\partial^2 \log f(y_{ij};\mu_{ij},\sigma^2)}{\partial \mu_{ij}^2}\Big],$$

it leads to

$$\mathrm{Var}(u_{ij}) = \frac{\sigma^2}{2}\mathrm{E}[d''(y_{ij};\mu_{ij})].$$

Define the pseudo response data by

$$\tilde{Y}_{ij} = \mu_{ij} + V_{ij}u_{ij}, \tag{4.37}$$

where $V_{ij} = \sigma^2/\mathrm{Var}(u_{ij})$. The mean of $\tilde{Y}_{ij}$ is $\mathrm{E}(\tilde{Y}_{ij}) = \mu_{ij}$ and the variance of $\tilde{Y}_{ij}$ is

$$\mathrm{Var}(\tilde{Y}_{ij}) = V_{ij}^2\mathrm{Var}(u_{ij}) = \sigma^2 V_{ij}.$$

Let $\tilde{\boldsymbol{Y}}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{it_i})'$ be the pseudo response vector and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{it_i})'$ be the mean vector for subject i. Assuming that the repeated observations are independent, then based on section 3.1.2 of previous chapter, the score function for the simplex distribution is

$$\boldsymbol{S}_I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\sigma^2} \boldsymbol{D}_i' \boldsymbol{u}_i = \sum_{i=1}^{n} \boldsymbol{D}_i' (\sigma^2 \boldsymbol{V}_i)^{-1} (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}_i), \tag{4.38}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ and $\boldsymbol{V}_i = \text{diag}[V_{i1}, \ldots, V_{it_i}]$.

In general cases, we need to take the correlation between the repeated measures on the same subject into account. Let $R(\boldsymbol{\alpha})$ be a working correlation matrix for $\tilde{\boldsymbol{Y}}_i$, where $\boldsymbol{\alpha}$ are the association parameters. The working covariance matrix for $\tilde{\boldsymbol{Y}}_i$ is specified by

$$\Sigma_i = \sigma^2 \boldsymbol{V}_i^{1/2} R(\boldsymbol{\alpha}) \boldsymbol{V}_i^{1/2}.$$

We define the generalized estimating equations for the simplex distribution to be

$$\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} (\tilde{\boldsymbol{y}}_i - \boldsymbol{\mu}_i). \tag{4.39}$$

When $R(\boldsymbol{\alpha}) = I$, $\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ reduces to the score functions for the independent responses $\boldsymbol{S}_I(\boldsymbol{\beta})$.

Let $\hat{\boldsymbol{\beta}}$ be the solutions to $\boldsymbol{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$. Arts and Jørgensen (2000) showed that under regularity conditions and certain conditions, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and asymptotically normal,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\boldsymbol{0}, \boldsymbol{V}) \qquad \text{as} \quad n \to \infty,$$

where the covariance matrix $\boldsymbol{V}$ is given by

$$\boldsymbol{V} = \lim_{n \to \infty} n I_0^{-1} I_1 I_0^{-1},$$

and

$$I_0 = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \boldsymbol{D}_i, \quad I_1 = \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \text{cov}(\tilde{\boldsymbol{Y}}_i) \Sigma_i^{-1} \boldsymbol{D}_i,$$

with $\text{cov}(\tilde{\boldsymbol{Y}}_i) = (\boldsymbol{V}_i \boldsymbol{u}_i)(\boldsymbol{V}_i \boldsymbol{u}_i)'$. The GEE estimate $\hat{\boldsymbol{\beta}}$ is computed by iterating between a modified Fisher scoring algorithm and estimation for $\boldsymbol{\alpha}$ and σ^2. First one needs to select a working correlation matrix. Given the current estimates of $\boldsymbol{\alpha}$

and σ^2, the approximate estimate for $\boldsymbol{\beta}$ is solved by the iteration

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \Big\{ \sum_{i=1}^{n} \boldsymbol{D}_i' \Sigma_i^{-1} \boldsymbol{D}_i \Big\}^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}). \qquad (4.40)$$

The robust variance estimator for $\hat{\boldsymbol{\beta}}$ is estimated by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \hat{I}_0^{-1} \hat{I}_1 \hat{I}_0^{-1}.$$

The dispersion parameter σ^2 and $\boldsymbol{\alpha}$ can be estimated by the method of moments. Based on the property (1) of the simplex distribution $\text{E}d(y;\mu) = \sigma^2$, σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{n} t_i - p} \sum_{i=1}^{n} \sum_{j=1}^{t_i} d(y_{ij}; \hat{\mu}_{ij}).$$

Since $\tilde{Y}_{ij} = \mu_{ij} + V_{ij} u_{ij}$, $\text{corr}(\tilde{Y}_{ij}, \tilde{Y}_{ij'}) = \text{corr}(u_{ij}, u_{ij'})$. Use the current estimation of $\hat{\boldsymbol{\beta}}$ to compute Pearson residuals for the pseudo responses

$$\hat{r}_{ij} = \frac{\tilde{y}_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{V}_{ij})}} = \frac{\hat{u}_{ij}}{\sqrt{(1/2)\text{E}d''(y_{ij}; \hat{\mu}_{ij})}}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, t_i.$$

For the exchangeable correlation, $\text{corr}(\tilde{Y}_{ij}, \tilde{Y}_{ij'}) = \alpha$ for all $j \neq j'$, α is estimated by

$$\hat{\alpha} = \frac{1}{\hat{\sigma}^2 \{ \sum_{i=1}^{n} \frac{1}{2} t_i (t_i - 1) - p \}} \sum_{i=1}^{n} \sum_{j' > j} \hat{r}_{ij'} \hat{r}_{ij}.$$

If one specifies an unstructured correlation matrix, $\boldsymbol{\alpha}$ is estimated by

$$\hat{\alpha}_{jj'} = \frac{1}{\hat{\sigma}^2 (n - p)} \sum_{i=1}^{n} \hat{r}_{ij} \hat{r}_{ij'}, \quad j, j' = 1, \ldots, t_i.$$

4.4.2 Mixtures of Marginal Models

In order to adapt the ME model into the correlated outcome data setting, Rosen, Jiang and Tanner (2000) proposed mixtures of marginal models. Their method is a combination of the ME model with the standard GEE method. In their paper, they assumed that the response data come from mixture distributions of the exponential family. Since the simplex distribution does not belong to the exponential family, we need to extend the mixtures of marginal models to repeated measures of compliance data. The method we present here combines the ME model we proposed in the previous chapter with the extension of the GEE method for the simplex distribution.

Assume for the response variable Y_{ij}, with probability π_{ij0}, $Y_{ij} = 0$; with probability π_{ijC}, $Y_{ij} = 1$; with probability π_{ijc}, $Y_{ij} \sim S_c^-(\mu_{ijc}, \sigma_c^2)$ $(c = 1, \ldots, C-1)$. Let $g(y_{ij}; \mu_{ijc}, \sigma_c^2)$ be the density function for simplex distribution $S^-(\mu_{ijc}, \sigma_c^2)$. The density function for response y_{ij} is:

$$\begin{aligned} f(y_{ij}) &= \pi_{ij0} I(y_{ij} = 0) + \sum_{c=1}^{C-1} \pi_{ijc} g(y_{ij}; \mu_{ijc}, \sigma_c^2) I(0 < y_{ij} < 1) + \pi_{ijC} I(y_{ij} = 1) \\ &= \sum_{c=0}^{C} \pi_{ijc} f_c(y_{ij}; \mu_{ijc}), \end{aligned}$$

where $f_0(y_{ij}; \mu_{ij0}) = I(y_{ij} = 0)$, $f_C(y_{ij}; \mu_{ijC}) = I(y_{ij} = 1)$, and

$$f_c(y_{ij}; \mu_{ijc}) = g(y_{ij}; \mu_{ijc}, \sigma_c^2) I(0 < y_{ij} < 1), \quad c = 1, \ldots, C-1.$$

We use a multinomial logit model to model the marginal weights π_{ijc} and logit models to model the marginal means of the simplex distributions. That is,

$$\log \frac{\pi_{ijc}}{\pi_{ij0}} = \boldsymbol{x}'_{ij} \boldsymbol{\gamma}_c, \qquad c = 1, \ldots, C, \tag{4.41}$$

and

$$\operatorname{logit}(\mu_{ijc}) = \boldsymbol{x}'_{ij} \boldsymbol{\beta}_c, \qquad c = 1, \ldots, C-1. \tag{4.42}$$

First assume that repeated observations on a subject are independent. We analyze the data as in the previous chapter. Let us denote d_{ijc} $(c = 0, \ldots, C)$ as an indicator that represents whether Y_{ij} is from the c^{th} component distribution, $\Pr(d_{ijc} = 1) = \pi_{ijc}$ and $\sum_{c=0}^{C} d_{ijc} = 1$. Since for $c = 1, \ldots, C-1$, d_{ijc} are unknown, we treat them as missing values. The parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ are estimated through the EM algorithm. The log-likelihood for the complete data is

$$\ell_{(c)}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{c=0}^{C} d_{ijc}[\log f_j(y_{ij};\boldsymbol{\beta}_j) + \log \pi_{ijc}(\boldsymbol{\gamma})].$$

At the $(t+1)^{th}$ E-step, we replace the missing data by their expectation values conditional on the observed data and the parameter values at the t^{th} step.

$$\begin{aligned} E[\ell_{(c)}(\boldsymbol{\psi}^{(t+1)}|\boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc}^{(t)}\big(\log f_c(y_{ij};\boldsymbol{\beta}_c) + \log \pi_{ijc}(\boldsymbol{\gamma})\big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc}^{(t)} \log \pi_{ijc}(\boldsymbol{\gamma}) + \sum_{c=1}^{C-1}\Big[\sum_{i=1}^{n}\sum_{j=1}^{t_i} w_{ijc}^{(t)} \log f_c(y_{ij};\boldsymbol{\beta}_c)\Big], \end{aligned}$$

where

$$w_{ijc}^{(t)} = \frac{\pi_{ijc}^{(t)} f_j(y_{ij};\boldsymbol{\beta}_c^{(t)})}{\sum_{l=0}^{C}\pi_{ijl}^{(t)} f_l(y_{ij};\boldsymbol{\beta}_l^{(t)})}, \quad c = 0,\ldots,C. \tag{4.43}$$

The $(t+1)^{th}$ M-Step is to maximize

$$\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc}^{(t)} \log \pi_{ijc}(\boldsymbol{\gamma}), \tag{4.44}$$

and

$$\max_{\boldsymbol{\beta}_c} \sum_{i=1}^{n}\sum_{j=1}^{t_i} w_{ijc}^{(t)} \log f_c(y_{ij};\boldsymbol{\beta}_c), \quad c = 1,\ldots,C-1. \tag{4.45}$$

The maximization with respect to $\boldsymbol{\gamma}$ is the solution to the estimating equation

$$S(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{c=0}^{C} w_{ijc}^{(t)} \frac{\partial \log \pi_{ijc}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = 0. \tag{4.46}$$

The maximization with respect to $\boldsymbol{\beta}_c$ is a weighted version of the ML estimator for a simplex distribution. For subject i, define $\boldsymbol{u}_{i,c} = (u_{i1c},\ldots,u_{it_ic})'$ to be the score vector, with $u_{ijc} = -(1/2)d'(y_{ij};\mu_{ijc})$; define $\tilde{\boldsymbol{Y}}_{i,c} = (\tilde{Y}_{i1c},\ldots,\tilde{Y}_{it_ic})'$ to be the pseudo response vector, where $\tilde{Y}_{ijc} = \mu_{ijc} + V_{ijc}u_{ijc}$ and $V_{ijc} = \sigma_c^2/\text{Var}(u_{ijc})$. From

Section 4.4.1, $\boldsymbol{\beta}_c^{(t)}$ are the solutions to the estimating equations:

$$\boldsymbol{S}(\boldsymbol{\beta}_c) = \sum_{i=1}^{n} \frac{1}{\sigma^2} \boldsymbol{D}'_{i,c} \boldsymbol{H}_{i,c} \boldsymbol{u}_{i,c} = \sum_{i=1}^{n} \boldsymbol{D}'_{i,c} (\sigma_c^2 \boldsymbol{V}_{i,c})^{-1} \boldsymbol{H}_{i,c} (\tilde{\boldsymbol{y}}_{i,c} - \boldsymbol{\mu}_{i,c}), \quad c = 1, \ldots, C-1, \tag{4.47}$$

where

$$\boldsymbol{D}_{i,c} = \partial \boldsymbol{\mu}_{i,c} / \partial \boldsymbol{\beta}_c, \qquad \boldsymbol{V}_{i,c} = \text{diag}[V_{i1c}, \ldots, V_{it_ic}],$$

and

$$\boldsymbol{H}_{i,c} = \text{diag}[w_{i1c}, \ldots, w_{it_ic}].$$

To account for dependence of repeated observations on a subject, we incorporate the extension of the GEE method for the simplex distribution into estimating equations (4.47). This results in a generalization of the EM algorithm. Rosen, Jiang and Tanner (2000) called it the Expectation-Solution (ES) algorithm. Let $R(\boldsymbol{\alpha}_c)$ be the working correlation matrix for the c^{th} component. Denote

$$\boldsymbol{\Sigma}_{i,c} = \sigma_c^2 \boldsymbol{V}_{i,c}^{1/2} R(\boldsymbol{\alpha}_c) \boldsymbol{V}_{i,c}^{1/2}.$$

The $(t+1)^{th}$ S-Step is to solve the estimating equations:

$$\boldsymbol{S}(\boldsymbol{\beta}_c, \boldsymbol{\alpha}_c) = \sum_{i=1}^{n} \boldsymbol{D}'_{i,c} \boldsymbol{\Sigma}_{i,c}^{-1} \boldsymbol{H}_{i,c} (\tilde{\boldsymbol{y}}_{i,c} - \boldsymbol{\mu}_{i,c}), \quad c = 1, \ldots, C-1. \tag{4.48}$$

Therefore, the S-Step is to solve a system of weighted extension of GEE method for each simplex model. The GEE estimate $\hat{\boldsymbol{\beta}}_c$ is computed by iterating between a modified Fisher scoring algorithm and estimation for $\boldsymbol{\alpha}_c$ and σ_c^2. First we select a working correlation matrix for each component. Given the current estimates of $\boldsymbol{\alpha}_c$ and σ_c^2, the approximate estimate for $\boldsymbol{\beta}_c$ is solved by the iteration

$$\boldsymbol{\beta}_c{}^{(t+1)} = \boldsymbol{\beta}_c^{(t)} + \left\{ \boldsymbol{D}_{i,c}^{(t)\prime} \boldsymbol{\Sigma}_{i,c}^{(t)^{-1}} \boldsymbol{H}_{i,c}^{(t)} \boldsymbol{D}_{i,c}^{(t)} \right\}^{-1} \boldsymbol{S}(\boldsymbol{\beta}_c^{(t)}, \boldsymbol{\alpha}_c^{(t)}), \quad c = 1, \ldots, C-1. \tag{4.49}$$

The dispersion parameter σ_c^2 and $\boldsymbol{\alpha}_c$ can be estimated by the method of moments, which is

$$\hat{\sigma}_c^2 = \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{t_i} w_{ijc} - p_c} \sum_{i=1}^{n} \sum_{j=1}^{t_i} w_{ijc} d(y_{ij}; \hat{\mu}_{ijc}) I(0 < y_{ij} < 1), \quad c = 1, \ldots, C-1,$$

where $p_c = \dim(\boldsymbol{\beta}_c)$. We use the current estimation of $\hat{\boldsymbol{\beta}}_c$ to compute Pearson residuals for the pseudo responses

$$\hat{r}_{ijc} = \frac{\tilde{y}_{ijc} - \hat{\mu}_{ijc}}{\sqrt{V(\hat{V}_{ijc})}} I(0 < y_{ij} < 1) = \frac{\hat{u}_{ijc} I(0 < y_{ij} < 1)}{\sqrt{(1/2)\mathrm{E}d''(y_{ij}; \hat{\mu}_{ijc})}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, t_i.$$

For the exchangeable correlation, α_c is estimated by

$$\hat{\alpha}_c = \frac{1}{\hat{\sigma}_c^2 \{ \sum_{i=1}^{n} \sum_{j'>j} w_{ijc} w_{ij'c} - p_c \}} \sum_{i=1}^{n} \sum_{j'>j} w_{ijc} w_{ij'c} \hat{r}_{ij'c} \hat{r}_{ijc}.$$

If one specifies an unstructured correlation matrix, $\boldsymbol{\alpha}_c$ is estimated by

$$\hat{\alpha}_{jj'c} = \frac{1}{\hat{\sigma}_c^2 (n-p)} \sum_{i=1}^{n} w_{ijc} w_{ij'c} \hat{r}_{ijc} \hat{r}_{ij'c}, \quad j, j' = 1, \ldots, t_i.$$

4.4.3 Expectation-Solution Algorithm

Rosen et al. (2000) noted since (4.48) is not the summand of the data score function, unlike the EM algorithm, the ES algorithm is not guaranteed to converge. However, if the algorithm does converge, it converges to a solution to an unbiased estimating equation, which is consistent and asymptotically normal. We show that our estimating equations satisfy the conditions needed for the Rosen et al. (2000) proposition to hold. Thus, if the ES algorithm for our compliance problems converges, it converges to a consistent and asymptotically normal estimator.

Rosen et al. (2000) considered a more general algorithm for solving estimating equations with incomplete data. Here we adapt their theory to the ES algorithm we proposed for our repeated compliance problems. We denote

$\boldsymbol{d}_{ij} = (d_{ij0}, \ldots, d_{ijC})'$ to be the indicator vector. Thus, $(\boldsymbol{d}_{ij}|\boldsymbol{\pi}_{ij})$ are *i.i.d.* with multinomial distribution $\prod_{c=0}^{C} \pi_{ijc}^{d_{ijc}}$. The joint distribution of $(y_{ij}, \boldsymbol{d}_{ij})$ has density

$$p(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) = \prod_{c=0}^{C} \left[\pi_{ijc}(\boldsymbol{\gamma}) f_c(y_{ij}; \boldsymbol{\beta}_c)\right]^{d_{ijc}}.$$

Let us define $q(\boldsymbol{y}, \boldsymbol{d}|; \boldsymbol{\psi})$ to be a $d_{\boldsymbol{\psi}} \times 1$ dimensional vector function ($d_{\boldsymbol{\psi}} = \dim(\boldsymbol{\psi})$), such that $q(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\psi})$ is measurable and integrable with respect to $p(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\psi})$, and $q(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\psi})$ is continuously differentiable on $\boldsymbol{\psi}$ for each $(y_{ij}, \boldsymbol{d}_{ij})$.

The Expectation-Solution algorithm is: at the $(t+1)^{th}$ iteration, the Expectation step is to compute

$$\begin{aligned} S(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)}) &= \sum_{i=1}^{n}\sum_{j=1}^{t_i} \mathrm{E}[q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|y_{ij}, \boldsymbol{\psi}^{(t)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{t_i} \int q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) p(\boldsymbol{d}_{ij}|y_{ij}, \boldsymbol{\psi}^{(t)}) d\boldsymbol{d}_{ij}. \end{aligned}$$

The Solutions step is solve for $\widehat{\boldsymbol{\psi}} = \boldsymbol{\psi}^{(t+1)}$ from the equation $S(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)}) = 0$.

Proposition (Rosen et al. 2000): Assume that the following conditions hold:

(a) $q(.,.;.)$ is an unbiased estimating function, satisfying

$$E[q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{\psi}] = \int q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{\psi}) p(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) dy_{ij} d\boldsymbol{d}_{ij} = 0$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, t_i$.

(b) $S(\boldsymbol{\psi}|\boldsymbol{\psi})$ is a continuous function on $\boldsymbol{\psi}$.

If there exists a point $\widehat{\boldsymbol{\psi}}$ such that $\lim_{t\to\infty} \boldsymbol{\psi}^{(t)} = \widehat{\boldsymbol{\psi}}$, where $\boldsymbol{\psi}^{(t)}$, $t = 0, 1, 2, \ldots$, is the solution for the t^{th} iteration of the ES algorithm, then:

(i) $\widehat{\boldsymbol{\psi}}$ satisfies the estimating equation $S(\widehat{\boldsymbol{\psi}}|\widehat{\boldsymbol{\psi}}) = 0$;

(ii) $S(\boldsymbol{\psi}|\boldsymbol{\psi}) = 0$ is an unbiased estimating equation, satisfying $E[S(\boldsymbol{\psi}|\boldsymbol{\psi})] = 0$.

Now we prove that the estimating equations in our ES algorithm satisfy the conditions of the Proposition. At the $(t+1)^{th}$ iteration, the conditional distribution of $\boldsymbol{d}_{ij}$ given y_{ij} is

$$p(d_{ijc} = 1|y_{ij}, \boldsymbol{\psi}^{(t)}) = w_{ijc}^{(t)} = \frac{\pi_{ijc}^{(t)} f_j(y_{ij}; \boldsymbol{\beta}_c^{(t)})}{\sum_{l=0}^{C} \pi_{ijl}^{(t)} f_l(y_{ij}; \boldsymbol{\beta}_l^{(t)})}, \quad c = 0, \ldots, C.$$

The estimation step becomes

$$S(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{c=0}^{C} q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) w_{ijc}^{(t)} = 0.$$

The first $d_{\boldsymbol{\gamma}} \times 1$ vector estimating equation is $S(\boldsymbol{\gamma})$, where $d_{\boldsymbol{\gamma}} = \dim(\boldsymbol{\gamma})$.

$$S(\boldsymbol{\gamma}|\boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{c=0}^{C} q_{\boldsymbol{\gamma}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) w_{ijc}^{(t)} = 0,$$

where

$$q_{\boldsymbol{\gamma}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) = \frac{\partial \log \pi_{ijc}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}.$$

The next $d_{\boldsymbol{\beta}} \times 1$ vector estimating equation is $S(\boldsymbol{\beta}, \boldsymbol{\alpha}|\boldsymbol{\psi}^{(t)})$, where $d_{\boldsymbol{\beta}} = \dim(\boldsymbol{\beta})$.

$$S(\boldsymbol{\beta}, \boldsymbol{\alpha}|\boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{c=0}^{C} q_{\boldsymbol{\beta}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) w_{ijc}^{(t)} = 0,$$

where $q_{\boldsymbol{\beta}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi}) = (q_{11}, \ldots, q_{1d_{\boldsymbol{\beta}_1}}, \ldots, q_{C-1,1}, \ldots, q_{C-1,d_{\boldsymbol{\beta}_{C-1}}})'$. The components of $q_{\boldsymbol{\beta}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})$ are $q_{rl} = [\boldsymbol{B}_{ijc}]_{rl}(\tilde{y}_{ijc} - \boldsymbol{\mu}_{ijc})$ for $r = 1, \ldots, C-1$ and $l = 1, \ldots, d_{\boldsymbol{\beta}_c}$, where $[\boldsymbol{B}_{ijc}]_{rl} = \sum_{u=1}^{t_i} [\boldsymbol{D}'_{i,c}]_{ul} [\Sigma_{i,c}^{-1}]_{uj} \delta_{cr}$ and δ_{cr} is Kronecker's delta. Here we use $[\boldsymbol{A}]_{rl}$ to represent the component at the r^{th} row and l^{th} column of the matrix $\boldsymbol{A}$. The estimating equation $S(\boldsymbol{\beta}, \boldsymbol{\alpha}|\boldsymbol{\psi}^{(t)})$ actually equals equation (4.47).

We can easily see that condition (b) of the Proposition holds. Now we need to show that $q_{\boldsymbol{\gamma}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})$ and $q_{\boldsymbol{\beta}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})$ satisfy condition (a) of the Proposition. From the conditional distribution of $(\boldsymbol{d}_{ij}|y_{ij}, \boldsymbol{\psi})$,

$$\mathrm{E}[q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{\psi}] = \sum_{c=0}^{C} w_{ijc} \mathrm{E}[q(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{d}_{ij}, \boldsymbol{\psi}].$$

For $q_{\boldsymbol{\gamma}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})$, condition (a) holds because

$$\begin{aligned}
\mathrm{E}[q_{\boldsymbol{\gamma}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{\psi}] &= \sum_{c=0}^{C} \pi_{ijc} \mathrm{E}\left[\frac{\partial \log \pi_{ijc}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\bigg|\boldsymbol{d}_{ij}, \boldsymbol{\psi}\right] \\
&= \sum_{c=0}^{C} \frac{\partial \pi_{ijc}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \sum_{c=0}^{C} \pi_{ijl}(\delta_{cl} - \pi_{ijc}) \\
&= \pi_{ijl}(1 - \sum_{c=0}^{C} \pi_{ijc}) = 0.
\end{aligned}$$

For $q_{\boldsymbol{\beta}}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})$, condition (a) holds because

$$\begin{aligned}
\mathrm{E}[q_{\boldsymbol{\beta}_c}(y_{ij}, \boldsymbol{d}_{ij}; \boldsymbol{\psi})|\boldsymbol{\psi}] &= \sum_{c=0}^{C} \pi_{ijc} \mathrm{E}\Big[[\boldsymbol{B}_{ijc}]_{rl}(\tilde{y}_{ijc} - \boldsymbol{\mu}_{ijc})|\boldsymbol{d}_{ij}, \boldsymbol{\psi}\Big] \\
&= \sum_{c=0}^{C} \pi_{ijc}[\boldsymbol{B}_{ijc}]_{rl}\Big[\mathrm{E}(\tilde{y}_{ijc}|\boldsymbol{d}_{ij}, \boldsymbol{\psi}) - \boldsymbol{\mu}_{ijc}\Big] = 0.
\end{aligned}$$

Carroll, Ruppert and Stefanski (1995) showed that solutions of unbiased estimating equations are consistent and asymptotically normal as $n \rightarrow \infty$. Denote the estimating equation $S(\boldsymbol{\psi}) = \sum_{i=1}^{n} S_i(\boldsymbol{\psi})$. The asymptotic variance of $\widehat{\boldsymbol{\psi}}$ can be estimated by $\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\psi}}) = \hat{I}_0^{-1}\hat{I}_1\{\hat{I}_0^{-1}\}'$, where $\hat{I}_0 = \frac{\partial S(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}|\widehat{\boldsymbol{\psi}}$ and $\hat{I}_1 = \sum_{i=1}^{n} S_i(\widehat{\boldsymbol{\psi}})S_i'(\widehat{\boldsymbol{\psi}})$.

4.5 Applications

4.5.1 Asthma Study I

This example is the same as the example we used in the previous chapter. In that study, we analyzed 90 children with persistent asthma who were taking one of the three medicines: inhaled corticosteroid (*ICS*), nebulized cromolyn (*CRO*), and slow–release theophylline (*SRT*). There were another 26 patients who were taking a combination of inhaled steroids and *SRT*. Each medicine had its own compliance rate. We did not incorporate them in the analysis of the cross-sectional study in the previous chapter. In this section, we analyze all 116 patients together. Table 4–1 shows some basic statistics for the three drugs.

Table 4–1: Basic information statistics for asthma study I

Medication	Total No.	Mean	Std. Dev.	No. of 0%	No. of 100%
ICS	69	0.624	0.341	6	17
CRO	14	0.449	0.378	2	3
SRT	59	0.721	0.265	0	10

As in the last chapter, a doctor gave an assessment to each patient's compliance. We call this explanatory variable *Assess* ($1 =$ very poor, $2 =$ less than optimal, and $3 =$ optimal). We also know each patient's age.

4.5.1.1 Subject-specific models

4.5.1.1.1 ME models with random effects

Before we fit the random effects ME model, we need to choose the number of components and the number of parameters. First we treated the data as independent observations and fitted them with the ME model we proposed in the previous chapter. To choose the number of components C in the ME model, we fitted the full ME models with different Cs. For observation y_{ij}, the gating (weight) network model is

$$\log \frac{\pi_{ijc}}{\pi_{ij0}} = \gamma_{c0} + \gamma_{c1}\text{ICS} + \gamma_{c2}\text{CRO} + \gamma_{c3}\text{Assess} + \gamma_{c4}\text{Age}, \quad c = 1, \ldots, C.$$

The components in the ME model are $C - 1$ simplex distributions. The mean of the c^{th} simplex distribution is modeled by a logit model,

$$\text{logit}(\mu_{ijc}) = \beta_{c0} + \beta_{c1}\text{ICS} + \beta_{c2}\text{CRO} + \beta_{c3}\text{Assess} + \beta_{c4}\text{Age}, \quad c = 1, \ldots, C - 1.$$

We fitted the ME models with $C = 2, 3, 4$, and used the model selection criteria AIC and AIC$_c$ to select the number of components. The results are given in Table 4–2. Since the sample size is small relative to the number of parameters ($n/d < 40$), the AIC criterion may under penalize the number of the parameters in the mixture. In this example, using AIC$_c$ is more appropriate. The AIC$_c$ value

Table 4–2: The criterion of selecting C

C	$\ell(\widehat{\boldsymbol{\psi}})$	No. of Para. d	AIC	AIC$_c$
2	-84.342	15	198.684	202.494
3	-56.611	25	163.622	174.829
4	-42.018	35	154.036	177.810

for $C = 3$ is the smallest one among these three ME models. Therefore, we chose $C = 3$ for this data set.

For $C = 3$, we performed the likelihood-ratio tests to choose the number of parameters. The final ME model includes

$$\log \frac{\pi_{ijc}}{\pi_{ij0}} = \gamma_{c0} + \gamma_{c3}\text{Assess}, \quad c = 1, 2, 3, \tag{4.50}$$

as the weights model;

$$\text{logit}(\mu_{ij1}) = \beta_{10} + \beta_{11}\text{ICS} + \beta_{12}\text{CRO} + \beta_{13}\text{Assess}, \tag{4.51}$$

as the model for the mean of the first simplex distribution component, and

$$\text{logit}(\mu_{ij2}) = \beta_{20}, \tag{4.52}$$

as the intercept model for the mean of the second simplex distribution component. This model has an estimated log-likelihood $\ell_1(\widehat{\boldsymbol{\psi}}) = -66.38$. In the random effects ME model with NPML model fitting, it is the case with $K = 1$ supporting point. With the same covariates, when we fitted the data with $K = 2$ supporting points, the log-likelihood is $\ell_2(\widehat{\boldsymbol{\psi}}) = -58.89$; with $K = 3$ supporting points, the log-likelihood $\ell_3(\widehat{\boldsymbol{\psi}}) = -58.49$. So the deviance difference $\text{dev}_2 = 14.98$. Compared to $\text{dev}_3 = 15.78$, the deviance difference does not change much. We chose $K = 2$ as the final model. Table 4–3 shows the NPML estimates of the final ME models.

From the ML estimates we can see that medicines had no significant effects on the weights. The estimated parameters $\hat{\gamma}_{13} > 0$, $\hat{\gamma}_{23} > 0$, $\hat{\gamma}_{33} > 0$ reveal that the

Table 4–3: Parameter estimation of the final model with $K = 1, 2$

Parameter	$K = 1$		$K = 2$	
	Estimate	S.E.	Estimate	S.E.
γ_{10} (Intercept)	0.147	0.857	0.613	0.884
γ_{13} (Assess)	1.232	0.535	1.147	0.539
γ_{20} (Intercept)	-1.600	1.029	-1.259	1.483
γ_{23} (Assess)	1.405	0.578	1.047	0.718
γ_{30} (Intercept)	-2.841	1.155	-2.383	1.194
γ_{33} (Assess)	2.023	0.602	1.893	0.611
β_{10} (Intercept)	-0.676	0.309	-0.363	0.284
β_{11} (ICS-SRT)	-0.438	0.209	-0.472	0.173
β_{12} (CRO-SRT)	-1.127	0.389	-1.072	0.311
$\beta_{11} - \beta_{12}$ (ICS-CRO)	0.688	0.389	0.600	0.309
β_{13} (Assess)	0.497	0.118	0.445	0.104
β_{20} (Intercept)	1.699	0.483	1.915	0.204

higher the assessments are, the bigger the weights are comparing to the weight at mass point 0. For patients coming from the first simplex distribution component, $\hat{\beta}_{11} = -0.472$ and $\hat{\beta}_{12} = -1.072$. These tell us that in this group of patients, using ICS and CRO resulted in lower compliances than using SRT. For patients coming from the second simplex distribution component, medicines and doctor's assessment have no significant effects on patients' compliances.

4.5.1.1.2 Cumulative logit model with random effects

As in the last chapter, we again used the same three grouping methods: Grouping method one – 3 categories (0% - 50%, 51% to 84% and 85% to 100%); grouping method two – 4 categories (0% – 25%, 26% – 50%, 51% – 75%, and 76% – 100%); grouping method three – 5 categories (0% – 20%, 21% – 40%, 41% – 60%, 61% – 80%, and 81% – 100%). We started with the full model with all covariates and interaction terms. We used the SAS procedure NLMIXED to fit the models. Through likelihood-ratio tests, the three methods gave the same final model

Table 4–4: Parameter estimation for 3-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	1.469	0.829	1.384	1.045
θ_2	3.309	1.063	2.976	2.122
β_1 (ICS-SRT)	-0.277	0.453	-0.220	0.425
β_2 (CRO-SRT)	-2.019	0.972	-1.574	1.698
$\beta_1 - \beta_2$ (ICS-CRO)	1.742	0.924		
β_3 (Assess)	1.191	0.384	1.056	0.764
γ_1 (ICS-SRT)			0.194	0.525
γ_2 (CRO-SRT)			-0.176	1.905
γ_3 (Assess)			-0.105	0.289
σ	1.570	0.793	1.382	1.168
$-2\ell(\widehat{\psi})$	280.8		280.5	

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - (\beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess} + b_i), \quad k = 1, \ldots, K-1, \quad (4.53)$$

where $b_i \sim N(0, \sigma^2)$ accounts for within-subject correlation.

Based on the final model, we also fitted a scaled cumulative logit model with random effects

$$\text{logit}[P(Y_{ij,g} \leq k)] = \frac{\theta_k - (\beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess} + b_i)}{\exp(\gamma_1\text{ICS} + \gamma_2\text{CRO} + \gamma_3\text{Assess})}, \quad k = 1, \ldots, K-1.$$

Tables 4–4 to 4–6 give the ML estimates for the three grouping methods.

Comparing the scaled random effects cumulative logit models with standard random effects cumulative logit models, we can see that the log-likelihoods have little changes. Therefore, the standard cumulative logit models with random effects are sufficient to fit these grouped data sets. All the final models show that the compliance to CRO was significantly lower than the compliance to ICS and SRT conditional on doctor's assessment. The compliance to ICS and the compliance to SRT did not have a significant difference. For example, for 5-category grouped responses, $\hat{\beta}_2 = -2.319$ with a standard error of 0.863. The

Table 4–5: Parameter estimation for 4-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	-0.143	0.905	-0.051	1.311
θ_2	1.941	0.922	3.111	2.471
θ_3	3.354	1.061	5.247	3.813
β_1 (ICS-SRT)	-0.603	0.554	-0.796	1.035
β_2 (CRO-SRT)	-2.975	1.142	-4.349	3.204
$\beta_1 - \beta_2$ (ICS-CRO)	2.372	1.029		
β_3 (Assess)	1.608	0.443	2.516	1.826
γ_1 (ICS-SRT)			0.296	0.573
γ_2 (CRO-SRT)			-2.309	1.825
γ_3 (Assess)			0.123	0.271
σ	2.126	0.824	3.295	2.472
-2$\ell(\widehat{\psi})$	312.3		311.2	

Table 4–6: Parameter estimation for 5-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	-0.517	0.771	-0.740	1.085
θ_2	1.320	0.766	1.643	1.491
θ_3	2.389	0.844	2.935	2.353
θ_4	3.350	0.944	4.060	3.165
β_1 (ICS-SRT)	-0.607	0.467	-0.486	0.637
β_2 (CRO-SRT)	-2.319	0.863	-2.556	2.233
$\beta_1 - \beta_2$ (ICS-CRO)	1.712	0.800		
β_3 (Assess)	1.450	0.371	1.655	1.296
γ_2 (ICS-SRT)			0.669	0.569
γ_2 (CRO-SRT)			0.681	0.794
γ_3 (Assess)			-0.105	0.210
σ	1.602	0.663	0.954	1.770
-2$\ell(\widehat{\psi})$	370.5		367.7	

estimated odds that a patient's compliance to medicine SRT falls below any fixed category are $\exp(-2.319) = 0.10$ times the estimated odds of his compliance to CRO conditional on doctor's assessment. The estimated odds that a patient's compliance to medicine CRO falls below any fixed category are $\exp(1.712) = 5.54$ ($\beta_1 - \beta_2 = 1.712$ with S.E. $= 0.800$) times the estimated odds of his compliance to ICS conditional on doctor's assessment. The doctor's assessment can be used to predict the patient compliance. For 5-category grouped responses, conditional on a given medicine, the estimated odds that a patient's compliance falls below any fixed category decrease by $\exp(1.450) = 4.26$ if the doctor's assessment increases by 1.

4.5.1.2 Population-averaged models

4.5.1.2.1 Standard GEE Method

With the GEE approach, we did not specify the underlying distribution for the observations. We just assumed that the marginal mean of the compliance variable Y_{ij} is μ_{ij}, and the marginal variance of Y_{ij} is $\text{Var}(Y_{ij}) = \phi V(\mu_{ij})$. We chose $V(\mu_{ij})$ to be (a) $\mu_{ij}(1-\mu_{ij})$ and (b) $[\mu_{ij}(1-\mu_{ij})]^2$. Since the maximum number of the repeated observations is 2, we chose the independence structure and the exchangeable correlation structure for the working correlation matrix.

We chose the final model through comparing the deviances of the nested models. For both variance functions, the final model is

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess}. \tag{4.54}$$

Note that the final model picks the same covariates as the final model in the cumulative logit model with random effects. Tables 4–7 and 4–8 give the GEE estimates for both variance functions.

These variance functions give similar parameter estimates. The models give $\hat{\beta}_2 < 0$ and $\hat{\beta}_1 - \hat{\beta}_2 < 0$. They are both significant, which imply that patients who

Table 4–7: GEE Parameter estimation for variance function $V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$

Parameter	Indep. Corr.		Exch. Corr.	
	Estimate	S.E.	Estimate	S.E.
β_0 (Intercept)	-0.819	0.355	-0.753	0.352
β_1 (ICS-SRT)	-0.253	0.227	-0.281	0.214
β_2 (CRO-SRT)	-1.181	0.362	-1.147	0.344
$\beta_1 - \beta_2$ (ICS-CRO)	0.928	0.368	0.866	0.356
β_3 (Assess)	0.748	0.139	0.717	0.136
ϕ	0.624		0.624	

Table 4–8: GEE Parameter estimation for variance function $V(\mu_{ij}) = [\mu_{ij}^2(1 - \mu_{ij})]^2$

Parameter	Indep. Corr.		Exch. Corr.	
	Estimate	S.E.	Estimate	S.E.
β_0 (Intercept)	-0.827	0.357	-0.745	0.353
β_1 (ICS-SRT)	-0.237	0.229	-0.281	0.216
β_2 (CRO-SRT)	-1.217	0.353	-1.182	0.333
$\beta_1 - \beta_2$ (ICS-CRO)	0.980	0.359	0.901	0.345
β_3 (Assess)	0.740	0.138	0.707	0.135
ϕ	1.395		1.395	

Table 4–9: GEE parameter estimation for grouped responses

Parameter	3 categories		4 categories		5 categories	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
θ_1	1.109	0.551	0.115	0.526	-0.260	0.538
θ_2	2.423	0.590	1.435	0.550	1.064	0.544
θ_3			2.294	0.568	1.823	0.553
θ_4					2.506	0.587
β_1 (ICS-SRT)	-0.143	0.324	-0.213	0.324	-0.334	0.317
β_2 (CRO-SRT)	-1.471	0.593	-1.866	0.554	-1.711	0.567
$\beta_1 - \beta_2$ (ICS-CRO)	1.328	0.584	1.652	0.566	1.377	0.559
β_3 (Assess)	0.872	0.201	1.059	0.210	1.062	0.207

were taking CRO had lower compliances than patients who were taking ICS or SRT conditional on the doctor's assessment. The compliance to ICS and the compliance to SRT did not have on significant difference conditional on doctor's assessment. The doctor's assessment can be used to predict a patient's compliance. The higher the assessment, the higher the mean compliance.

4.5.1.2.2 GEE method for grouped data

We used the GEE method to fit marginal models for the grouped data sets. The grouping methods were the same as the ones in the random effects cumulative logit model fitting . The final models for the grouped data are

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - (\beta_1\text{ICS} + \beta_2\text{CRO} + \beta_3\text{Assess}), \quad k = 1, \ldots, K-1. \quad (4.55)$$

where $K = 3, 4, 5$.

We used the SAS procedure GENMOD to get the GEE estimates of the marginal cumulative logit model with independent correlation structure. The fitting results are given in Table 4–9. Compared with the estimated effects in the random effects cumulative logit models, the GEE estimated effects in the marginal model are smaller. This is because that the marginal effects average the heterogeneity of the conditional effects. Agresti (2002, pp. 499–501) gave a detailed

Table 4–10: Parameter estimation for mixtures of marginal models with exchangeable correlation structure

Parameter	Estimate	S.E.
γ_{10} (Intercept)	0.150	0.858
γ_{13} (Assess)	1.236	0.537
γ_{20} (Intercept)	-1.626	0.939
γ_{23} (Assess)	1.395	0.557
γ_{30} (Intercept)	-2.841	1.243
γ_{33} (Assess)	2.023	0.631
β_{10} (Intercept)	-0.622	0.278
β_{11} (ICS-SRT)	-0.509	0.258
β_{12} (CRO-SRT)	-1.057	0.292
$\beta_{11} - \beta_{12}$ (ICS-CRO)	0.549	0.199
β_{13} (Assess)	0.476	0.105
β_{20} (Intercept)	1.794	0.155

explanation of why the marginal effects are smaller than the conditional effects. But the conclusions are the same as the ones in the random effects models.

4.5.1.2.3 Mixture of marginal models

As in the random effects ME model, we first treated the responses as independent and fit the regular ME models. We used the model selection methods for cross-sectional data to choose the number of components and the number of parameters. This is the same as the first step in fitting the NPML estimates in the random effects ME model. The final ME model includes models 4.50–4.52. We chose the exchangeable correlation structure as the working correlation matrix. The estimated parameters with their standard errors are given in Table 4–10. The parameter estimates are close to the ML estimates in the cross-sectional study (Table 4–3, $K = 1$). The explanation of the estimated parameters are similar to the one in the random effects ME model section.

For the first simplex component, the exchangeable correlation parameter is estimated as 0.72. For the second simplex component, the exchangeable correlation parameter is estimated as -0.31. Therefore, in the first distribution component,

Table 4–11: The estimated mean compliances conditional on assessment

Mean	Assess=1	Assess=2	Assess=3
ICS (S.E.)	0.399 (0.052)	0.580 (0.063)	0.734 (0.071)
CRO (S.E.)	0.328 (0.055)	0.496 (0.072)	0.660 (0.085)
SRT (S.E.)	0.477 (0.055)	0.662 (0.056)	0.798 (0.057)

Table 4–12: The estimated compliance differences conditional on assessment

Difference	Assess =1	Assess = 2	Assess = 3
SRT − ICS (S.E.)	0.078 (0.040)	0.082 (0.042)	0.064 (0.035)
SRT − CRO (S.E.)	0.149 (0.041)	0.165 (0.047)	0.139 (0.047)
ICS − CRO (S.E.)	0.071 (0.037)	0.084 (0.045)	0.074 (0.042)

the repeated measures are positively correlated, which means if a patient's compliance to the first medicine was high, his compliance to the second medicine was likely to be high. This represents the group of people that are consistent in their compliances to the medicines. In the second distribution component, the repeated measures are negatively correlated, which means if a patient's compliance to the first medicine was high, his compliance to the second medicine was likely to be low. This represents the group of people that prefer one medicine to the other medicine.

Table 4–11 gives the estimated population mean compliance of each drug conditional on the assessment. Table 4–12 gives the estimated differences of the compliances conditional on the assessment. From Table 4–12, we can see that patients who were taking SRT always had higher predicted compliances than patients who were taking CRO. At assessment level 1 and 2, patients who were taking SRT had higher predicted compliances than patients who were taking ICS at significance level 0.05.

4.5.2 Asthma Study II

The second example is also an asthma medicines study (Sherman et al. 2001). The purpose of this study is to evaluate compliance to oral montelukast (OM) and inhaled fluticasone (IF) in children with persistent asthma. Surveys have

Table 4–13: Basic information statistics for asthma study II

Medication	No. of Obs.	Mean Comp.	No. of 0%	No. of 100%	Mean Time	Mean Age
OM	123	0.579	5	7	228.260	8.361
IF	117	0.466	8	6	278.564	8.206

shown that adolescent patients prefer oral to inhaled agents (Sherman et al. 2001). Therefore, it is expected that compliance rates to OM are higher than compliance rates to IF. The total patients studied were 170 children with persistent asthma. A combination of both medicines was prescribed for 69 children. Each of the medicine had its own compliance rate. OM alone was prescribed for 54 children and IF alone was prescribed for 48 children. The compliance was calculated as (no. of doses refilled)/(no. of doses prescribed)×100%. This is not the actual compliance rate, but the maximum possible compliance rate.

The covariates include the observation period (*Time*), the prescribed drug (*OM* or *IF*) and patients' age (*Age*). All patients except two had been on the prescription for at least 90 days. The average observation period was 228 days for OM and 279 days for IF. The mean age of the patients was 8.3 years. Some basic statistics of the data set are given in Table 4–13.

4.5.2.1 Subject-specific model

4.5.2.1.1 ME models with random effects

The first step is to choose the number of components and the number of parameters. We treated the responses as independent observations and fit the full ME models with different numbers of components. We used drug IF as the baseline drug. We treated covariate OM as a dummy variable. The full ME models have the form of

$$\log \frac{\pi_{ijc}}{\pi_{ij0}} = \gamma_{c0} + \gamma_{c1}\mathrm{OM} + \gamma_{c2}\mathrm{Time} + \gamma_{c3}\mathrm{Age}, \quad c = 1, \ldots, C,$$

Table 4–14: The criterion of selecting C

C	$\ell(\widehat{\psi})$	No. of Para. d	AIC	AIC$_c$
2	-102.329	12	228.658	230.032
3	-69.897	20	179.794	183.630
4	-24.838	28	105.676	113.373

as the gating (weight) network model, and

$$\text{logit}(\mu_{ijc}) = \beta_{c0} + \beta_{c1}\text{OM} + \beta_{c2}\text{Time} + \beta_{c3}\text{Age}, \quad c = 1, \ldots, C-1,$$

as the logit models for the means of the simplex distribution components.

We fitted the ME models with $C = 2, 3, 4$, and used the model selection criteria AIC and AIC$_c$ to select the number of components. The results are given in Table 4–14. Both criteria imply that we should choose $C = 4$. But when we took a close look at the fitted ME model for $C = 4$, we found that one simplex component has an estimated dispersion parameter $\hat{\sigma}^2 = 1.5e - 8$. This means that patients in this sub-population had compliances clumping at a mass point between 0 and 1. From the density function of a simplex distribution

$$f(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma^2}\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2} \Big\},$$

we can see that for a patient who belonged to this component group and had a compliance close to the mean, the probability $f(y_{ij}; \mu, \sigma^2) \approx [2\pi\sigma^2\{y_{ij}(1-y_{ij})\}^3]^{-1/2}$ would be extremely high. The density function for this subject would be

$$f(y_i; \boldsymbol{\theta}) = \pi_{i0} I(y_i = 0) + \sum_{c=1}^{3} \pi_{ic} f_c(y_i; \mu_{ic}) I(0 < y_i < 1) + \pi_{i4} I(y_i = 1).$$

If we treat it as a clumping point instead of following a simplex distribution. The density function would be

$$\begin{aligned} f(y_i;\boldsymbol{\theta}) &= \pi_{i0}I(y_i=0)+\sum_{c=1}^{2}\pi_{ic}f_c(y_i;\mu_{ic})I(0<y_i<1) \\ &\quad +\pi_{i3}I(y_i=\mu_{i3})+\pi_{i4}I(y_i=1). \end{aligned}$$

This gives a much smaller density value. Therefore, these few observations result in a much bigger maximum log-likelihood compared to the maximum log-likelihood for $C=3$. Choosing $C=4$ is not so appropriate. Therefore, we chose $C=3$ instead of $C=4$.

For $C=3$, we performed the likelihood-ratio tests to choose the number of parameters. The final ME model includes the weight model

$$\log\frac{\pi_{ijc}}{\pi_{ij0}}=\gamma_{c0},\quad c=1,3, \tag{4.56}$$

for the first component and the last component, and a weight model

$$\log\frac{\pi_{ij2}}{\pi_{ij0}}=\gamma_{20}+\gamma_{21}\text{OM}+\gamma_{22}\text{Time}, \tag{4.57}$$

for the second component. The model for the mean of the first simplex distribution component is

$$\text{logit}(\mu_{ij1})=\beta_{10}+\beta_{12}\text{Time}. \tag{4.58}$$

The model for the mean of the second simplex distribution component is

$$\text{logit}(\mu_{ij2})=\beta_{20}+\beta_{21}\text{OM}. \tag{4.59}$$

In the NPML analysis, the above model is a nonparametric random effects model with $K=1$ support point. This model has $-2\ell(\widehat{\boldsymbol{\psi}})=149.83$. We fitted the data with $K=2$ supporting points, $-2\ell(\widehat{\boldsymbol{\psi}})=139.46$; with $K=3$ supporting points, $-2\ell(\widehat{\boldsymbol{\psi}})=137.58$. So the deviance difference $\text{dev}_2=10.37$. Compared to

Table 4–15: Parameter estimation of the final model with $K = 1, 2$

Parameter	$K = 1$		$K = 2$	
	Estimate	S.E.	Estimate	S.E.
γ_{10} (Intercept)	2.676	0.283	2.518	0.300
γ_{20} (Intercept)	-4.920	1.235	-5.398	1.742
γ_{21} (OM-IF)	1.789	0.450	1.725	0.876
γ_{22} (Time)	0.015	0.004	0.017	0.005
γ_{30} (Intercept)	0	0.392	-0.018	0.382
β_{10} (Intercept)	5.876	1.035	6.597	0.665
β_{12} (Time)	-0.015	0.003	-0.017	0.002
β_{20} (Intercept)	-0.191	0.093	-0.233	0.100
β_{21} (OM-IF)	0.496	0.133	0.480	0.136

$\text{dev}_3 = 12.25$, the deviance difference does not change much. We chose $K = 2$ as the final model. Table 4–15 shows the NPML estimates of the final ME model.

Based on the NPML estimates, the first component weight was not affected by any covariates. Since $\hat{\beta}_{12} = -0.017$, for patients belonging to this component group, the longer they stayed with the prescribed drug, the lower the compliance rates were. The estimated compliances to OM and IF have no significant difference in the first component group. The estimated parameter $\hat{\gamma}_{21} = 1.725$ implies that patients taking OM had higher probabilities of belonging to the second component group than patients taking IF. The longer the observation period, the more likely the patient belonged to the second component group ($\hat{\gamma}_{22} = 0.017$). In the second component group, patients taking OM had higher compliance rates than patients taking IF. The estimated parameter for the third weight model is close to 0, which implies that the estimated weight of having a 100% compliance was similar to the estimated weight of having a 0% compliance for both medicines.

Table 4–16: Parameter estimation 3-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	0.453	0.222	0.443	0.227
θ_2	2.429	0.362	2.492	0.450
β_1 (OM-IF)	0.858	0.287	0.855	0.296
γ_1 (OM-IF)			0.068	0.259
σ	0.904	0.390	0.923	0.418
-$2\ell(\widehat{\psi})$	471.4		471.3	

4.5.2.1.2 Cumulative logit model with random effects

We used the same grouping methods as in the first example. We started with the full model

$$\begin{aligned}\text{logit}[P(Y_{ij,g} \leq k)] &= \theta_k - (\beta_1\text{OM} + \beta_2\text{Time} + \beta_3\text{Age} + \beta_4\text{OM} * \text{Time} \\ &\quad +\beta_5\text{OM} * \text{Age} + \beta_5\text{Time} * \text{Age} + b_i), \quad k = 1, \ldots, K-1,\end{aligned}$$

where $b_i \sim N(0, \sigma^2)$ accounts for the within-subject correlation. We used SAS NLMIXED to fit the models. All the three grouping methods chose the same final model

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - (\beta_1\text{OM} + b_i), \quad k = 1, \ldots, K-1. \tag{4.60}$$

We also fitted a scaled cumulative logit model with random effects

$$\text{logit}[P(Y_{ij,g} \leq k)] = \frac{\theta_k - (\beta_1\text{OM} + b_i)}{\exp(\gamma_1\text{OM})}, \quad k = 1, \ldots, K-1.$$

The ML estimates for the three grouping methods are given in Tables 4–16 to 4–18.

For all three grouped data sets, the added scaled parameters do not change the log-likelihood too much. Therefore, the standard cumulative logit models with random effects are sufficient to fit the three grouped data sets. All the final models show that the compliance to OM was significantly higher than the compliance to IF. The observation period and age had no significant effects on patients'

Table 4–17: Parameter estimation for 4-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	-1.292	0.272	-1.450	0.379
θ_2	0.565	0.246	0.608	0.273
θ_3	1.895	0.335	2.099	0.484
β_1 (OM-IF)	1.024	0.297	1.149	0.384
γ_1 (OM-IF)			0.201	0.261
σ	1.366	0.385	1.533	0.519
$-2\ell(\widehat{\psi})$	643.0		642.3	

Table 4–18: Parameter estimation for 5-category grouped responses

Parameter	Final Model		Scaled Model	
	Estimate	S.E.	Estimate	S.E.
θ_1	-1.874	0.312	-2.194	0.490
θ_2	-0.104	0.233	-0.153	0.260
θ_3	1.073	0.268	1.224	0.344
θ_4	2.417	0.367	2.825	0.606
β_1 (OM-IF)	1.003	0.286	1.181	0.387
γ_1 (OM-IF)			0.291	0.242
σ	1.453	0.357	1.710	0.528
$-2\ell(\widehat{\psi})$	747.1		745.5	

Table 4–19: GEE Parameter estimation for variance function $V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$

Parameter	Indep. Corr.		Exch. Corr.	
	Estimate	S.E.	Estimate	S.E.
β_0 (Intercept)	-0.137	0.103	-0.150	0.102
β_1 (OM-IF)	0.454	0.134	0.447	0.129
ϕ	0.570		0.570	

Table 4–20: GEE Parameter estimation for variance function $V(\mu_{ij}) = [\mu_{ij}(1-\mu_{ij})]^2$

Parameter	Indep. Corr.		Exch. Corr.	
	Estimate	S.E.	Estimate	S.E.
β_0 (Intercept)	-0.137	0.103	-0.150	0.102
β_1 (OM-IF)	0.454	0.134	0.447	0.129
ϕ	1.400		1.149	

compliances. For grouped responses using method one, $\hat{\beta}_1 = 0.858$ with a standard error of 0.287. The estimated odds that a patient's compliance to medicine IF falls below any fixed category are exp(0.858) = 2.358 times the estimated odds of his compliance to OM.

4.5.2.2 Population-averaged models

4.5.2.2.1 Standard GEE method

As in the first example, we chose the variance function $V(\mu_{ij})$ to be (a) $\mu_{ij}(1 - \mu_{ij})$ and (b) $[\mu_{ij}(1 - \mu_{ij})]^2$. Since the maximum number of the repeated observations is 2, we chose the independence structure and the exchangeable correlation structure for the working correlation matrix.

We chose the final model through comparing $2(Q(\hat{\boldsymbol{\beta}}_1; \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}_0; \boldsymbol{y}))$ for nested models. For both variance functions, the final model is

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \text{OM}. \tag{4.61}$$

Tables 4–19 and 4–20 give the GEE estimates for both variance functions.

Table 4–21: GEE parameter estimation for grouped responses

Parameter	3 groups		4 groups		5 groups	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
θ_1	0.377	0.185	-1.025	0.192	-1.448	0.205
θ_2	2.094	0.234	0.372	0.178	-0.118	0.171
θ_3			1.374	0.198	0.730	0.179
θ_4					1.725	0.204
β_1 (OM-IF)	0.740	0.250	0.741	0.236	0.714	0.210

The two variance functions give exactly the same parameter estimates. For the exchangeable correlation, the estimated parameter of OM is 0.447 with S.E. 0.129. Therefore, the compliance to OM was significantly higher than the compliance to IF.

4.5.2.2.2 GEE method for grouped data

We used the GEE method to fit marginal models for the grouped data sets. The final models for the grouped data are

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - \beta_1\text{OM}, \quad k = 1, \ldots, K-1. \tag{4.62}$$

where $K = 3, 4, 5$.

We used the SAS procedure GENMOD to get the GEE estimates with an independent correlation structure. Table 4–21 gives the GEE parameter estimates for the grouped responses. The estimated drug effect are close in the three grouped data sets. For grouping method one, $\hat{\beta}_1 = 0.740$. The estimated odds that a patient's compliance to medicine IF falls below any fixed category are $\exp(0.740) = 2.10$ times the estimated odds of his compliance to OM.

4.5.2.2.3 Mixture of marginal models

The first step is the same as in the random effects ME model. Therefore, the final model includes models 4.56–4.59 . We chose the exchangeable correlation

Table 4–22: Parameter estimation for mixtures of marginal model with exchangeable correlation structure

Parameter	Estimate	S.E.
γ_{10} (Intercept)	2.677	0.288
γ_{20} (Intercept)	-4.936	1.251
γ_{21} (OM-IF)	1.792	0.304
γ_{22} (Time)	0.015	0.003
γ_{30} (Intercept)	0	0.392
β_{10} (Intercept)	5.892	0.319
β_{12}(Time)	-0.015	0.001
β_{20} (Intercept)	-0.190	0.082
β_{21} (OM-IF)	0.497	0.111

structure as the working correlation matrix. The estimated parameters with their standard errors are given in Table 4–22.

For the first simplex component, the exchangeable correlation parameter is estimated as -0.194. For the second simplex component, the exchangeable correlation parameter is estimated as 0.130. The correlation parameters are relatively small, which implies that the repeated measures on the same subject were not highly correlated. Conditional on the mean observation period, the estimated mean compliance to IF is 47.02% and the estimated mean compliance to OM is 59.56%. The estimated difference between the mean compliances to IF and the mean compliance to OM is -12.53% with a standard error of 0.026. Therefore, the mean compliance to OM was significantly higher than the compliance to IF.

CHAPTER 5
CONCLUSIONS

5.1 Summary

Data with clumps occur in applications in many areas. In this dissertation, we developed methods for modeling a few special cases of this type of data, including repeated measures of zero-inflated count data, cross-sectional compliance data, and repeated measures of compliance data. For the analysis of repeated measures of zero-inflated count data, we proposed a correlated random effects hurdle model, a zero-altered mixed model and a random effects cumulative logit model. For the analysis of cross-sectional compliance data, we presented a two-part model, a mixtures of experts model, a cumulative logit model, and a quasi-likelihood method. Then, we extended the methods we proposed in the cross-sectional compliance data analysis into the repeated measures setting, where we considered both the subject-specific model approach and the population-averaged model approach.

To apply the methods we proposed, we used three real-life examples in biometrics for illustration. The first example is an occupational injury prevention program study, which was used to illustrate the random effects models for repeated measures of zero-inflated count data. The second and the third examples are both asthma medication studies for children with persistent asthma. Part of the second data set was used for illustration of models for cross-sectional compliance data. Both the second data set and third data set were used for illustration of the random effects models and the marginal models for repeated measures of compliance data.

In Chapter 2 we provided a detailed introduction of the hurdle model, including the technical details for the ML estimation for a Poisson hurdle model and a negative binomial hurdle model. Then we introduced a special case of the hurdle model called a zero-altered model, which could be used to test the existence of zero-inflation for the count data. We illustrated that the zero-altered model also has a parsimony property and is easy to be used to summarize covariate effects overall. We compared the hurdle model with the zero-inflated count model and we gave the reasons why we preferred the hurdle model to the zero-inflated count model. We conducted two small simulation studies to discuss the advantages of the hurdle model over the zero-inflated count model. The hurdle model and the zero-altered model were extended to include random effects for fitting repeated measures of zero-inflated count data. In maximum likelihood model fitting, we considered both a multivariate normal distribution and a nonparametric approach for the distribution of the correlated random effects. For ML model fitting with normal random effects, we proposed Gauss-Hermite quadrature to approximate the marginal log-likelihood and used an approximate version of the Fisher scoring method to obtain the ML estimates. For ML model fitting with a nonparametric approach, we applied the EM algorithm to obtain the NPML estimates. In addition to the NPML parameter estimation, we estimated the standard errors using Louis' approximation and discussed how to choose the number of mass points.

We also introduced a random effects cumulative logit model for analyzing repeated measures of zero-inflated count data. We considered a multivariate normal distribution for the random effects and the ML fitting was similar to the ML model fitting for the hurdle model with multivariate normal random effects. At the end of this chapter, we studied the identifiability of the ZIP model and used two simulation experiments to study the identifiability of the random effects ZIP model.

Some results from this chapter are as follows. The random effects hurdle model is easier to fit than the random effects zero-inflated count model. It is not only suitable for zero-inflation problems, but also for modeling data with fewer zeros than would be expected under standard distribution assumptions. When a data set is zero-deflated at some levels of the covariates, the zero-inflated count model may fail. The hurdle model does not have this problem. With the simple random intercepts form of the model, one can use the the existing software SAS NLMIXED to fit the correlated random effects hurdle model. The zero-altered mixed model can be used to test the existence of the zero-inflation for correlated count data. A simpler form of the zero-altered mixed model has the convenient property of summarizing the covariate effects overall. If one groups the possible count outcomes into ordered categories, a random effects cumulative logit model can be used. This model has the simplicity of using a single model to handle the clump at 0 and the positive outcomes. This is an important advantage over the random effects hurdle model, which needs to average results from both parts of the model together. The disadvantage of the random effects cumulative logit model is that it models the grouped data instead of the original data. The fixed effects ZIP model is identifiable. Simulation studies showed that the random effects ZIP model may be also identifiable.

Chapter 3 and Chapter 4 focused on compliance data study, which has two boundary clumps. Chapter 3 studied modeling cross-sectional compliance data. We first introduced a two-part model to analyze compliance data. The first part models the discrete masses at 0 and 1, and the second part models the continuous proportions between 0 and 1. For the first part of the model we proposed the cumulative logit model and the multinomial logit model. Both of them can be fitted using SAS procedures. For continuous proportions between 0 and 1, we introduced the logistic-normal distribution, the beta distribution, and the simplex distribution. Since the simplex distribution shares some common analytic properties with the exponential family, we used it in the rest of the analysis. We then generalized the two-part model to an ME model. This ME model can handle the case when the response variable follows a bimodal or multimodal distribution. A multinomial logit model was proposed to fit the weights of the ME model and logit models were proposed to fit the means of the simplex components. In model fitting, we implemented the EM algorithm and used Louis' approximation to approximate the standard error estimation. In addition, we discussed using AIC and AIC_c criteria to select the number of components and using likelihood-ratio tests to select the number of parameters. We implemented the delta method in comparing the average compliances of several groups.

The ME model can reveal the underlying sub-populations and find out the different covariate effects on different sub-populations. However, when comparing the overall mean compliances of different groups that are levels of the explanatory variables, one needs to average results from all components to make an unconditional comparison. To make this kind of inference easier, we proposed two single-model approaches to handle the extreme values (zeros and ones) and the continuous proportional responses together. The first approach is the ordinal threshold model approach. We discussed the score test to check the proportional odds assumption. If the assumption does not hold, we suggested applying the scaled cumulative logit model. The second approach we proposed is the quasi-likelihood method. Besides discussing the model fitting and the hypothesis test for nested models, we also proposed a quasi-score test for choosing between the logit link function and the complementary log-log link function.

In Chapter 4, we extended the methods we proposed in Chapter 3 for repeated measures of compliance data. Since the subject-specific model approach and the population-averaged model approach are useful in different situations, we considered both types of approaches in modeling repeated measures of compliance data. At the subject level, we proposed a random effects ME model and briefly mentioned a random effects scaled cumulative logit model. For the random effects ME model, we used a nonparametric assumption for the distribution of random effects because of the convenience of model fitting. We applied the EM algorithm to obtain the NPML estimation and Louis' method to approximate the complicated standard error estimation. At the population level, we first introduced the standard GEE method and the GEE method for repeated ordinal response data. Our emphasis was on the mixtures of marginal models, which is the marginal ME model. We extended the GEE method to the simplex distribution. Then we combined this extension of the GEE method with the ME model to form the mixtures of marginal models. A generalization of the EM algorithm, the ES algorithm, was introduced to fit the mixtures of marginal models. We also discussed using the method of moments to calculate the dispersion parameters and the correlation parameters.

For both the cross-sectional study and the repeated measures study for compliance data, when one wants to compare the mean compliances of several groups, the single-model approaches are preferred. They are easier to fit and easier to use in making comparisons. When one wants to study the covariate effects on the compliances of the underlying sub-populations, the ME model is preferred.

5.2 Future Research

There are a few interesting areas in this dissertation that are possible topics for future research. For semicontinuous data analysis, most of the methods we reviewed assume that the positive continuous responses have a log-normal distribution. This is not necessarily realistic, especially in applications in which some especially large observations create a right tail that is too long for a log-normal distribution. For instance, in a survey of medical care expenses, the right tail may be poorly modeled by the log-normal distribution. Semi-parametric methods may be appropriate for such highly skewed data. For repeated measures of semicontinuous data, there are a few papers on applying random effects models at the subject level. However, there is little attention on using the marginal models for population-averaged study. This is a potential area for future research. The exponential dispersion model with $V(\mu) = \mu^p$ $(1 < p < 2)$ (Jørgensen 1987, 1997) has the simplicity of using a single model to analyze semicontinuous data. It may be of interest to extend this type of model to longitudinal data analysis. So far this method has essentially been ignored even for cross-sectional analysis.

The ordinal threshold model can be used in analyzing all three types of data with clumps that we discussed in this dissertation. This model has the simplicity of using a single model to handle the clumps and the rest of the data. However, before applying the ordinal threshold model, we need to group the possible outcomes into ordered categories. There are many open questions with regard to the grouping methods, such as how to choose the number of categories, how to choose the cutpoints, and in what ways different grouping methods give consistent conclusions. These questions could use more work in future study.

In model selection for the random effects model with the NPML approach (in Chapter 2 and Chapter 4), we choose the number of mass points K by increasing it from 1 until the change in the deviances is small. There is no formal significance test to use because of the nonstandard situation (the simpler model is on the boundary of the parameter space). Therefore, future research is needed to develop a formal test for this problem.

The ME model is a finite mixture model with the weights depending on covariates. For mixtures of distributions, identifiability is an important issue. Jiang and Tanner (1999) considered the identifiability of a certain type of ME model, in which the components ("experts") are from the exponential family. They showed the conditions for identifiability to hold are validated for Poisson, gamma, normal and binomial experts. In the ME model we proposed, the main components are the simplex distributions, which do not belong to the exponential family. It is important to extend the Jiang and Tanner (1999) approach to show that under certain conditions our proposed ME model is identifiable.

In the mixtures of marginal models, we only incorporated dependence structure into the expert network models. In fact, the weights of repeated observations on the same subject also tend to be correlated. If we could know the underlying sub-populations, we could use the GEE method for nominal categorical responses (Lipsitz et al. 1994) to model the correlated weights. However, we cannot observe to which sub-population the response belongs. More research is needed in incorporating the dependence structure into the multinomial logit model for the weights.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition, New York: Wiley.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses, *Biometrika* **67**: 261–272.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics* **55**: 117–128.

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models, *Journal of the Royal Statistical Society, Series B, Methodological* **47**: 67–75.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principles, Pages 267–281 *in* Petrov, B. N. and Csaki, F. (eds.) *Second International Symposium on Information Theory,* Akademiai Kiado, Budapest.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal, *Econometrica* **41**: 997–1016.

Amemiya, T. (1984). Tobit models: A survey, *Journal of Econometrics* **24**: 3–61.

Artes, R. and Jørgensen, B. (2000). Longitudinal data estimating equations for dispersion models, *Scandinavian Journal of Statistics* **27**: 321–334.

Arulampalam, W. and Booth, A. (1997). Who gets over the training hurdle? A study of the training experiences of young men and women in Britain, *Journal of Population Economics* **10**: 197–217.

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex, *Journal of Multivariate Analysis* **39**: 106–116.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology, *Journal of the Royal Statistical Society* **Supplement** 4: 137–183.

Böhning, D., Dietz, E. and Schlattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. *in* Rost, J. and Langeheine, R. (eds) *Applications of Latent Trait and Latent Class Models in the Social Sciences,* Waxmann: Münster.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika* **82**: 81–91.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*, New York: Springer-Verlag.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*, New York: Cambridge University Press.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, New York: Chapman and Hall.

Chang, B.-H. and Pocock, S. (2000). Analyzing data with clumping at zero – an example demonstration, *Journal of Clinical Epidemiology* **53**: 1036–1043.

Chen, K., Xu, L. and Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification, *Neural Networks* **12**: 1229–1252.

Clepper, I. (1992). Noncompliance, the invisible epidemic, *Drug Topics* **17**: 44–65.

Cowles, M. K., Carlin, B. P. and Connett, J. E. (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness, *Journal of the American Statistical Association* **91**: 86–98.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica* **39**: 829–844.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach, *Journal of Applied Econometrics* **12**: 313–336.

Detry, J. M. R., Block, P., Debacker, G., Degaute, J. P. and Six, R. (1995). Patient compliance and therapeutic coverage – comparison of amlodipine and slow-release nifedipine in the treatment of hypertension, *European Journal of Clinical Pharmacology* **47**: 477–481.

Dobbie, M. and Welsh, A. (2001a). Modelling correlated zero-inflated count data, *The Australian and New Zealand Journal of Statistics* **43**: 431–444.

Dobbie, M. and Welsh, A. (2001b). Models for zero-inflated count data using the Neyman type A distribution, *Statistical Modelling* **1**: 65–80.

Duan, N., Manning, Willard G., J., Morris, C. N. and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care (corr: V2 p413), *Journal of Business and Economic Statistics* **1**: 115–126.

Duan, N., Manning, Willard G., J., Morris, C. N. and Newhouse, J. P. (1984). Choosing between the sample-selection model and the multi-part model, *Journal of Business and Economic Statistics* **2**: 283–289.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edition, New York: Springer-Verlag.

Gerdtham, U. and Trivedi, P. (2001). Equity in swedish health care reconsidered: New results based on the finite mixture model, *Health Economics* **10**: 565–572.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **46**: 149–192.

Grogger, J. T. and Carson, R. T. (1991). Models for truncated counts, *Journal of Applied Econometrics* **6**: 225–238.

Gronau, R. (1974). Wage comparisons – a selectivity bias, *Journal of Political Economy* **82**: 1119–1144.

Grytten, J., Holst, D. and Laake, P. (1993). Accessibility of dental services according to family income in a non-insured population, *Social Science and Medicine* **37**: 1501–1508.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model, *Journal of Business and Economic Statistics* **9**: 215–222.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization, *Journal of Applied Econometrics* **12**: 225–242.

Gurmu, S. and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips, *Journal of Business and Economic Statistics* **14**: 469–477.

Hajivassiliou, V. A. (1994). A simulation estimation analysis of the external debt crises of developing countries, *Journal of Applied Econometrics* **9**: 109–131.

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study, *Biometrics* **56**: 1030–1039.

Hartzel, J., Agresti, A. and Caffo, B. (2001). Multinomial logit random effects models, *Statistical Modelling* **1**: 81–102.

Haynes, R. B., Taylor, D. W. and Sackett, D. L. (1979). *Compliance in Health Care*, Baltimore, Maryland: Johns Hopkins University Press.

Heckman, J. (1974). Shadow prices, market wages, and labor supply, *Econometrica* **42**: 679–694.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica* **47**: 153–161.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros, *Biometrical Journal. Journal of Mathematical Methods in Biosciences* **36**: 531–547.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**: 297–307.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts, *Neural Computation* **3**: 79–87.

Jansakul, N. and Hinde, J. (2002). Score tests for zero-inflated Poisson models, *Computational Statistics and Data Analysis* **40**: 75–96.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models, *Journal of the American Statistical Association* **93**: 720–729.

Jiang, W. and Tanner, M. (1999). On the identifiability of mixtures-of-experts, *Neural Networks* **12**: 1253–1258.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtres of experts and the EM algorithm, *Neural Computation* **11**: 181–214.

Jørgensen, B. (1987). Exponential dispersion models, *Journal of the Royal Statistical Society, Series B, Methodological* **49**: 127–145.

Jørgensen, B. (1997). *The Theory of Dispersion Models*, New York: Chapman and Hall.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics* **22**: 79–86.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics* **34**: 1–14.

Leung, S. F. and Yu, S. (1996). On the choice between sample selection and two-part models, *Journal of Econometrics* **72**: 197–229.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13–22.

Lipsitz, S. R., Kim, K. and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine* **13**: 1149–1163.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature, *Biometrika* **81**: 624–629.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series B, Methodological* **44**: 226–233.

Manning, W. G., Duan, N. and Rogers, W. H. (1987). Monte Carlo evidence on the choice between sample selection and two-part models, *Journal of Econometrics* **35**: 59–82.

McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society, Series B, Methodological* **42**: 109–142.

McCullagh, P. (1983). Quasi-likelihood functions, *The Annals of Statistics* **11**: 59–67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition, New York: Chapman and Hall.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*, New York: John Wiley and Sons.

Mclachlan, G. and Peel, D. (2001). *Finite Mixture Models*, New York: Wiley.

Melikian, C., White, T. J., Vanderplas, A., Dezii, C. M. and Chang, E. (2002). Adherence to oral antidiabetic therapy in a managed care organization: A comparison of monotherapy, combination therapy, and fixed-dose combination therapy, *Clinical Therapeutics* **24**: 460–467.

Melkersson, M. and Rooth, D.-O. (2000). Modeling female fertility using inflated count data models, *Journal of Population Economics* **13**: 189–203.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometrics* **33**: 341–365.

Olsen, M. and Schafer, J. (2001). A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association* **96**: 730–745.

Olsen, R. (1975). The analysis of two-variable models when one of the variables is dichotomous, *Yale University, Economics Dept. unpublished manuscript.*

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables, *Applied Statistics* **39**: 205–217.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**: 12–35.

Pohlmeier, W. and Ulrich, V. (1995). An econometric model of the two–part decision making process in the demand of health care, *Journal of Human Resources* **30**: 339–361.

Powell, J. L. (1986). Symmetrically trimmed least squares estimation for Tobit models, *Econometrica* **54**: 1435–1460.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**: 1033–1048.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation, *Proceedings of the Cambrudge Philosophical Society* **44**: 50–57.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edition, New York: Wiley.

Raudenbush, S. W., Yang, M.-l. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order Laplace approximation, *Journal of Computational and Graphical Statistics* **9**: 141–157.

Ridout, M., Hinde, J. and Demetrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, *Biometrics* **57**: 219–223.

Robinson, P. M. (1982). On the asymptotic properties of estimators of models containing limited dependent variables, *Econometrica* **50**: 27–42.

Rosen, O., Jiang, W. and Tanner, M. A. (2000). Mixtures of marginal models, *Biometrika* **87**(2): 391–404.

Saei, A., Ward, J. and McGilchrist, C. A. (1996). Threshold models in a methadone programme evaluation, *Statistics in Medicine* **15**: 2253–2260.

Shankar, V., Milton, J. and Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry, *Accident Analysis and Prevention* **29**: 829–837.

Sherman, J., Hutson, A., Baumstein, S. and Hendeles, L. (2000). Telephoning the patient's pharmacy to assess adherence with asthma medications by measuring refill rate for prescriptions, *Journal of Pediatrics* **132**: 532–536.

Sherman, J., Patel, P., Hutson, A., Chesrown, S. and Hendeles, L. (2001). Adherence to oral montelukast and inhaled fluticasone in children with persistent asthma, *Pharmacotherapy* **21**: 1464–1467.

Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data, *Biometrics* **56**: 496–502.

Teicher, H. (1961). Identifiability of mixtures, *Annals of Mathematical Statistics* **32**: 244–248.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica* **26**: 24–36.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families, *Statistics Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, Indian Statistical Institute (Calcutta), pp. 579–604.

van de Ven, W. and van Praag, B. (1981). The demand for deductibles in private health insurance: A probit model with sample selection, *Journal of Econometrics* **17**: 229–252.

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution, *Biometrics* **21**: 738–743.

Wang, K., Yau, K. K. and Lee, A. H. (2002). A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays, *Computer Methods and Programs in Biomedicine* **68**: 195–203.

Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates, *Biometrics* **52**: 381–400.

Waterhouse, D. M., Calzone, K. A., Mele, C. and Brenner, D. E. (1993). Adherence to oral tamoxifen – a comparison of patient self-report, pill counts, and microelectronic monitoring, *Journal of Clinical Oncology* **11**: 1189–1197.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**: 439–447.

Wedel, M., DeSarbo, W. S., Bult, J. R. and Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data, *Journal of Applied Econometrics* **8**: 397–411.

Wolfe, J. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Naval Personnel and Training Research Laboratory, *Technical Bulletin* **STB 72–2**: San Diego, California, USA.

Wu, J.-W. and Lee, W.-C. (2001). The quasi-score statistic in quasi-likelihood model, *Statistics* **35**: 523–535.

Yau, K. K. and Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme, *Statistics in Medicine* **20**: 2907–2920.

Yoo, S., Kim, T. and Lee, J. (2001). Modeling zero response data from willingness to pay surveys – a semi-parametric estimation, *Economics Letters* **71**: 191–196.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121–130.

BIOGRAPHICAL SKETCH

Yongyi Min was born on October 11th, 1974, in Changzhi, Shanxi province, China. She lived in Changzhi until 1987 and then moved to Xuzhou, Jiangsu province, with her parents and brother. In 1992, Yongyi went to Beijing to attend the Renmin University of China, where she studied in the Department of Statistics. After graduating with a Bachelor of Economics degree in 1996, she began her graduate study in the same department.

Before Yongyi finished her master's degree in China, she was accepted as a Ph.D. student in the Department of Statistics at the University of Florida. In the fall of 1998, she moved to Gainesville, Florida, and continued her graduate study in the US. During the first year and one half in the Department of Statistics, she worked as a teaching assistant. For the last four years, she has worked as a research assistant under her advisor, Dr. Alan Agresti. She obtained her Master of Statistics degree in August of 2000 and plans to receive her Ph.D. degree in December 2003. After graduating, Yongyi will move to New York City, where she has accepted a position with the United Nations.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Alan Agrest.

Alan Agresti, Chair
Distinguished Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James G. Booth

James Booth
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Ram Littell

Ramon Littell
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Rongling Wu
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina

James Algina
Professor of Educational Psychology

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 2003

Dean, Graduate School